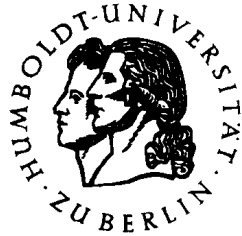


HUMBOLDT-UNIVERSITÄT ZU BERLIN
INSTITUT FÜR BIBLIOTHEKSWISSENSCHAFT



BERLINER HANDREICHUNGEN
ZUR BIBLIOTHEKSWISSENSCHAFT

HEFT 111

**ONTOLOGIEN IN DIGITALEN BIBLIOTHEKEN
UNTER DEM SCHWERPUNKT
INHALTSERSCHLIESSUNG UND RECHERCHE**

VON
SABINE BOLTZENDAHL

**ONTOLOGIEN IN DIGITALEN BIBLIOTHEKEN
UNTER DEM SCHWERPUNKT
INHALTSERSCHLIESSUNG UND RECHERCHE**

**VON
SABINE BOLTZENDAHL**

Berliner Handreichungen
zur Bibliothekswissenschaft

Begründet von Peter Zahn
Herausgegeben von
Konrad Umlauf
Humboldt-Universität zu Berlin

Heft 111

Boltzendahl, Sabine

Ontologien in digitalen Bibliotheken unter dem Schwerpunkt Inhaltserschließung und Recherche / von Sabine Boltzendahl. - Berlin : Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin, 2003, 42 S. - (Berliner Handreichungen zur Bibliothekswissenschaft und Bibliothekarsausbildung ; 111)

ISSN 14 38-76 62

Abstract:

Das Thema dieser Arbeit sind formale Ontologien in digitalen Bibliotheken. Diese Entwicklung der Künstlichen Intelligenz repräsentiert die Begrifflichkeit eines Fachgebietes oder thematischen Bereiches sowohl auf konzeptueller wie auf terminologischer Ebene, die durch vielfältige Beziehungen zwischen den Termen und integrierten Schlussregeln „intelligentes“ Suchen in Textdokumenten ermöglichen sollen. Ein Einsatz in digitalen Bibliotheken zur Bewältigung von Anforderungen des Information Retrieval und der Interoperabilität ist daher nahe liegend. Die vorliegende Arbeit betrachtet die Nutzung von Ontologien für Recherche und Inhaltserschließung in digitalen Bibliotheken und setzt sie in Beziehung zu den herkömmlichen dokumentarischen Hilfsmitteln Klassifikation und Thesaurus. Es zeigt sich, dass die Zielsetzungen bei der Entwicklung von formalen Ontologien nicht weit von den Anforderungen an moderne, elektronische Thesauri entfernt sind.

Diese Veröffentlichung entspricht einer schriftlichen Hausarbeit im Rahmen des postgradualen Fernstudiums Master of Arts (Library and Information Science) am Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin.

Inhalt

1	EINLEITUNG	6
2	FORMALE ONTOLOGIEN.....	7
2.1	WAS SIND FORMALE ONTOLOGIEN?	7
2.2	ANWENDUNGSBEREICHE VON ONTOLOGIEN	8
2.3	ARTEN VON ONTOLOGIEN	9
2.4	AUFBAU VON ONTOLOGIEN	10
2.5	BEISPIELE	12
2.5.1	<i>Cyc</i>	12
2.5.2	<i>WordNet</i>	13
2.5.3	<i>UMLS</i>	14
2.6	FAZIT	15
3	ONTOLOGIEN IN DIGITALEN BIBLIOTHEKEN	16
3.1	DIGITALE BIBLIOTHEKEN	16
3.1.1	<i>Information Retrieval</i>	17
3.1.2	<i>XML</i>	18
3.2	EINSATZ VON ONTOLOGIEN	19
3.2.1	<i>Recherche</i>	20
3.2.2	<i>Erschließung</i>	22
3.2.3	<i>Einbettung in die Systemarchitektur</i>	23
3.3	FAZIT	24
4	THESAURI UND ONTOLOGIEN – ALTER WEIN IN NEUEN SCHLÄUCHEN?.....	26
4.1	THESAURI – AUFBAU UND NUTZUNG.....	27
4.1.1	<i>Strukturelle Ebene</i>	27
4.1.2	<i>Terminologische Ebene</i>	29
4.1.3	<i>Nutzung</i>	29
4.2	DER NUTZEN VON ONTOLOGIEN.....	30
4.2.1	<i>Strukturelle Ebene</i>	31
4.2.2	<i>Terminologische Ebene</i>	33
4.2.3	<i>Nutzung</i>	34
4.3	FAZIT	35
5	FAZIT	37
6	ABKÜRZUNGSVERZEICHNIS.....	39
7	LITERATUR	40

1 Einleitung

Die vorliegende Arbeit befasst sich mit Anwendungsmöglichkeiten von „formalen Ontologien“ in digitalen Bibliotheken. Der Schwerpunkt liegt dabei auf der Inhaltsererschließung und der Recherche, also dem Information Retrieval.

Ontologien sind neuere Entwicklungen der künstlichen Intelligenz, die es sich zur Aufgabe gemacht haben, die Semantik elektronisch gespeicherter Daten zu beschreiben. Sie versprechen, damit angesichts einer steigenden Informationsflut, weltweiter Vernetzung und einem auch wirtschaftlich immer wichtiger werdenden Zugang zu und Austausch von Information und Wissen intelligente Lösungen zu liefern. Durch die Darstellung von gegenstandsbezogenem Wissen sollen mit Hilfe von Ontologien semantische Heterogenität behandelt und integriert, intelligente Recherchemöglichkeiten auch für große Datenmengen eröffnet und damit Lösungen v.a. für das Wissensmanagement geboten werden.

In dieser Arbeit soll der Frage nachgegangen werden, was unter Ontologien in diesem Sinne zu verstehen ist, wie sie in digitalen Bibliotheken als neuen Informationssystemen eingesetzt werden können und was sie im Verhältnis zu herkömmlichen Hilfsmitteln im Information Retrieval leisten.

Dafür soll zunächst erläutert werden, was formale Ontologien sind und wo und zu welchen Zwecken sie eingesetzt werden können (Kapitel 2). Im darauf folgenden Kapitel sollen zunächst Anforderungen an das Information Retrieval in digitalen Bibliotheken skizziert werden und mögliche Einsatzbereiche von Ontologien aufgezeigt werden (Kapitel 3). Dieses Kapitel wird durch Beispiele von konkreten Projekten, die Ontologien in digitalen Bibliotheken einsetzen, ergänzt. Schließlich sollen in Kapitel 4 Ontologien in Beziehung gesetzt werden zu herkömmlichen dokumentarischen Hilfsmitteln der Inhaltsererschließung und Recherche, v.a. dem Thesaurus.

Bei dieser Darstellung kann es sich nur um einen allgemeinen Überblick über den Themenbereich handeln, eine genauere Darstellung würde den Rahmen dieser Arbeit sprengen. Auf einige Probleme und mögliche Untersuchungsbereiche, die sich für den Einsatz von Ontologien in digitalen Bibliotheken stellen, soll jedoch noch am Schluss dieser Arbeit (Kapitel 5) hingewiesen werden. Auf besondere Ansätze z.B. in der Thesaurusforschung (wie z.B. einem „semiotischen Thesaurus“, vgl. Schwarz/Umstätter) kann in diesem Rahmen nicht eingegangen werden.

2 Formale Ontologien

Formale Ontologien werden seit den 1980er-, v.a. den 1990er-Jahren in der künstlichen Intelligenzforschung entwickelt. Das Ziel – allgemein formuliert – ist es, Maschinen zum „Verstehen“ des von ihnen gespeicherten Wissens zu bringen. Eine Anknüpfung an den philosophischen Begriff „Ontologie“ ergibt sich aus dem von manchen Vertretern der formalen Ontologie-Entwicklung erhobenen Anspruch, die „Welt“ bzw. das (gesamte) Wissen darüber in Form von eben diesen formalen Ontologien in Rechnern zu repräsentieren – ein Anspruch, der zum Teil recht vermessend erscheint. Fasst man einerseits Ontologie als „theory about the nature of being or the kinds existent“ (Fridman Noy/Hafner:53) und sieht andererseits die Aufgabe intelligenter Rechner-Systeme darin, eben diese Theorie über die Welt formal zu repräsentieren, so lässt sich eine Motivation für diese Benennung erkennen.

2.1 Was sind formale Ontologien?

Eine Ontologie ist eine Anordnung von (definierten) Konzepten und Relationen zwischen ihnen. Sie sind häufig hierarchisch organisiert und verfügen über Schlussmechanismen, die es ihnen erlauben, Aussagen über mit ihnen klassifizierte Gegenstände zu machen.

Eine einfache und vielfach zitierte¹ Definition von „Ontologie“ stammt von Gruber (1993):

„An ontology is a formal, explicit specification of a shared conceptualisation.“

Als Konzeptualisierung wird dabei eine Anordnung von Gegenständen, Konzepten und anderen Dingen, über die Wissen ausgedrückt wird, und die Beziehungen zwischen ihnen verstanden (vgl. Fridman Noy/Hafner 1997:53), diese Konzeptualisierung stellt ein abstraktes Modell und gleichzeitig eine bestimmte Sicht auf die Welt dar (vgl. Fensel:11; Vickery:278). Besondere Betonung liegt dabei darauf, dass es sich um geteiltes Wissen (und Terminologie) handelt, also um konsensuales, von den Mitgliedern einer Gruppe oder auch von Softwareagenten geteiltes Wissen (vgl. Ding:378; Fensel:1, 11; Uschold/Gruninger:96). Der Zugang zu diesem Wissen erfolgt dabei über Sprache: Ontologien bestehen aus einem bestimmten Vokabular, aus Definitionen ihrer Einträge und Relationen zwischen ihnen. Sie werden mit werden mit Wörterbüchern oder Glossaren verglichen, allerdings verfügen sie über reichere Strukturierungen zwischen und innerhalb der Einträge (vgl. Endres-Niggemeyer; SUO). Ontologien stellen demnach also zunächst ein (strukturiertes) Vokabular dar (vgl. Chandrasekaran u.a.:20; Fensel:11; Uschold/Gruninger:97), dessen Einträge auf die eine oder andere Weise inhaltlich spezifiziert sind.

Hier unterscheiden sich allerdings die Auffassungen und Darstellungsweisen von Ontologien: Ob es sich bei Ontologien nun um Vokabulare oder gar Wörterbücher handelt, oder aber ob es sich ausschließlich um Darstellungen von Konzepten und ihren Relationen handelt, ist strittig. Von manchen Autoren werden sie klar als terminologische Instrumente betrachtet (z.B. Endres-Niggemeyer; Vickery:278), andere Autoren sehen mehr die Konzepte, Relationen und Schlussmechanismen im Vorder-

¹ Vgl. z.B. Ding:378; Fensel:11; Fridman Noy/Hafner:53; Köpcke 2002:327; Vickery:278.

grund, die um eine bestimmte Terminologie ergänzt werden müssen (z.B. Amann u.a.: 226).

Die Spezifikation (auf diese bezieht sich das „explizit“ in der oben genannten Definition von Ontologie) bezieht sich jeweils auf das dahinter liegende, als eigentlich interessant betrachtete und mit dem Wort bezeichnete Konzept. Das, was hier strukturiert, angeordnet und zueinander in Relation gesetzt werden soll, sind die Konzepte, nicht die Wörter. „Spezifikation der Konzeptualisierung“ bedeutet die Definition des Konzeptes und seiner Eigenschaften, die Verdeutlichung von Relationen zwischen den Konzepten und die Erläuterung der Beschränkungen, die bei der Anwendung der Konzepte zu berücksichtigen sind. Durch diese Spezifikation soll implizites, verstecktes Wissen explizit und zugänglich gemacht werden. Dass die hier betrachteten Ontologien „formal“ sind, bedeutet im Wesentlichen, dass sie für Maschinen verarbeitbar sind (vgl. Fensel:11).

Ontologien stellen innerhalb der künstlichen Intelligenz (KI) so genannte „content theories“ dar, die sich – im Gegensatz zu den „mechanism theories“ – bei der Entwicklung von Expertensystemen weniger auf Schlussmechanismen und Ableitungsregeln konzentrieren, sondern sich vielmehr der Darstellung und Repräsentation von Inhalt, eben „content“, d.h. dem Sach- und Faktenwissen widmen (vgl. Chandrasekaran u.a.:21, 25; Fensel:1).²

2.2 Anwendungsbereiche von Ontologien

Als Zielsetzung und Einsatzbereich von Ontologien werden in der Literatur zunächst ganz allgemein fast alle Arten von Informationssystemen, die auf Wissen zurückgreifen, genannt. Sie sollen Anwendung finden in:³

- Wissensbasierten Systemen: Hier stellen sie Repräsentationen des jeweiligen Wissens dar (vgl. Chandrasekaran u.a.:24).
- Systementwicklung: Hier sollen sie eine genaue und standardisierte Modellierung des Anwendungsgebietes der zu entwickelnden Software liefern, zielgenaue Aufgabenspezifizierungen vornehmen, die Reliabilität und Konsistenz von Systemen und ggf. auch ihre Wiederverwendbarkeit in ähnlichen Zusammenhängen prüfen (vgl. Uschold/Gruninger:100ff). Sie sollen ein genaues Modell der im System zu repräsentierenden Gegenstände liefern (vgl. Chandrasekaran u.a.:23; Fensel:11).
- Natural Language Processing: Hier können sie durch Bereitstellung von Hintergrundwissen zur Disambiguierung bei der Interpretation von Wörtern beitragen⁴, zum anderen eine Rolle als „Wörterbuch“ von Konzepten spielen (vgl. Chandrasekaran u.a.:23; SUO).
- Wirtschaft: Ontologien sollen in der Kommunikation zwischen Menschen und Organisationen von Nutzen sein. Sie können (in firmeninterner Kommunikation) normativ wirken, indem sie verschiedene Perspektiven z.B. von verschiedenen Hierarchieebenen innerhalb eines Systems vereinheitlichen, für terminologische Konsistenz und Disambiguierung sorgen und Organisationsmodelle liefern. Damit

² Ausnahmen gibt es natürlich auch hier (vgl. Chandrasekaran u.a.:25); zudem werden gemeinsam mit dem Sachwissen auch Schlussregeln und Constraints in der Anwendung von Kategorien etc. mit abgelegt, aber es geht eben nicht um die Entwicklung von Schlussregeln unabhängig von Inhalten.

³ Vgl. zu dieser Unterteilung auch Köpcke 2002:328.

⁴ Als Beispiele könnten hier der Einsatz von Ontologien im automatischen Abstracting bzw. Zusammenfassen (vgl. Endres-Niggemeier) oder im automatischen Übersetzen (darauf weist Vickery (284) hin) dienen.

soll die Kommunikation verbessert werden, sowohl zwischen Menschen (z.B. Wissenschaftlern, die aus verschiedenen Disziplinen kommend denselben Gegenstand beforschen) als auch zwischen Software-Systemen (vgl. Uschold/Gruninger:98f). Den Daten- und Informationsaustausch sieht Fensel gerade für den E-Commerce als wichtiges Anwendungsgebiet.

- Information Retrieval und Wissensmanagement: Hier können Ontologien, z.B. in digitalen Bibliotheken, bei der Integration heterogener Informationsquellen und (Internet-)Suchmaschinen als Kategoriensysteme eingesetzt werden (vgl. Chandrasekaran u.a.:23). Die Ermöglichung von Daten- und Informationsaustausch auch zwischen verschiedenen Softwaresystemen auf der Basis von XML-Standards sieht Fensel als das zentrale Einsatzgebiet von Ontologien (vgl. Fensel:2ff), die als standardisierte Terminologie oder als Hilfsmittel zur Übersetzung zwischen verschiedenen Datenstrukturen dienen sollen (Fensel:8f). Dies soll der intelligenten Informationssuche und -extraktion, der Bereitstellung und Verwaltung von Textquellen und der Personalisierung von Websites dienen.

Die zentralen Aufgaben von Ontologien sind die Repräsentation (eines Ausschnittes) der Welt bzw. des Wissens darüber in einem abstrakten Modell sowie die Integration verschiedener Perspektiven und Terminologien sowohl zwischen Menschen als auch zwischen Softwaresystemen. Sie stellen konzeptuelle und terminologische Strukturen für verschiedene Gegenstandsbereiche zur Verfügung. Als terminologische Hilfsmittel finden sie Eingang in Instrumente des Natural Language Understanding.

Die Unterstützung von Interoperabilität von verschiedenen Software- oder Agentensystemen, die im Falle von wissensbasierten Systemen auf z.T. verschiedene Modelle von der Welt zurückgreifen, ist ein Bereich, in dem Ontologien im Information Retrieval sowohl in der Wirtschaft als auch in anderen Bereichen Anwendung finden sollen. Hier können Ontologien entweder als Hilfe zur Standardisierung von Terminologien der verschiedenen Datenbestände dienen oder als Mittel, mit Hilfe dessen Übersetzungen zwischen diesen Terminologien vorgenommen werden können (vgl. Uschold/Gruninger:99f; Fensel:8f).

Insgesamt lassen sich in den oftmals noch hypothetischen oder angestrebten Anwendungsbereichen von formalen Ontologien zwei Richtungen unterscheiden: zum einen Anwendungen, die darauf zielen, bessere oder neuere intelligente Software zu entwickeln (hier wäre z.B. die Integration von Ontologien in problemlösende Software anzusiedeln), zum anderen die Verwendung von Ontologien in Bereichen, in denen sie die Anwendung von Software verbessern sollen (und die dann natürlich eine Integration von Ontologien in die entsprechende Software voraussetzt): E-Commerce, Simulation und Modellierung von Organisationsabläufen und -systemen und v.a. Wissensmanagement und Information Retrieval, was hier von besonderem Interesse ist.

2.3 Arten von Ontologien

Die verschiedenen existierenden Ontologien lassen sich nach verschiedenen Kriterien sortieren.

Eine häufig wieder aufgenommene Unterteilung stammt von Uschold und Gruninger: Sie unterscheiden Ontologien nach dem Formalitätsgrad, in dem das

Vokabular und die dazugehörige Bedeutungs- oder Konzeptbeschreibung ausgedrückt ist. Sie nennen als zufällige Punkte auf einem angenommenen Kontinuum:

- sehr informal, d.h. der Ausdruck erfolgt ungefähr und in natürlicher Sprache;
- semi-informal, der Ausdruck erfolgt hier ebenfalls in natürlicher Sprache, diese ist aber kontrolliert, strukturiert und disambiguiert;
- semi-formal, hier erfolgt der Ausdruck in formal definierter, künstlicher Sprache;
- sehr formal, hier sind die Terme der formalen Semantik akribisch definiert und es gibt Sätze und Beweise für Begründetheit und Vollständigkeit (vgl. Uschold/Gruninger:97; Ding:378; Vickery:278).⁵

Ontologien lassen sich auch nach Anwendungsgebieten oder ihrem Allgemeingrad gruppieren. Fensel (vgl. Fensel:12) unterscheidet

- Bereichsontologien, die Wissen aus bestimmten Bereichen (Medizin, Mechanik usw.) beinhalten
- Metadatenontologien, z.B. Dublin-Core-Standards,
- allgemeine oder „Commonsense“-Ontologien, in denen allgemeines Wissen über die Welt niedergelegt ist und die daher in verschiedenen Bereichen einsetzbar sind,
- repräsentationale Ontologien, in denen verzeichnet ist, wie und mit welchen Kategorien Wissen wiederum in Ontologien dargestellt werden kann,
- Methoden- und Aufgaben-Ontologien, in denen spezifische Terminologie zur Aufgaben- oder Problemlösung abgelegt ist, die in problemlösender Software benötigt wird.

Darüber hinaus können noch linguistische Ontologien wie WordNet oder Ontolingua unterschieden werden, die besonderen Wert auf die Beschreibung der Terminologie legen (vgl. z.B. Weinstein/Birmingham:22).

2.4 Aufbau von Ontologien

Der Aufbau von Ontologien (wenn im Einzelnen auch recht verschieden, s.u.) erfolgt im Allgemeinen in einer Taxonomie, die zumeist durch IS-A-Relationen hierarchisch strukturiert ist (vgl. Fridman Noy/Hafner:55; Chandrasekaran u.a.:22; Weinstein/Birmingham:22).⁶ Die genaue Ausgestaltung dieser Taxonomie variiert von Ontologie zu Ontologie, ihre Existenz wird – v.a. bei so genannten Upper-Level-Ontologien und je nach abgebildetem Bereich – als nicht unbedingt nötig angesehen. (vgl. Fridman Noy/Hafner:68). Innerhalb dieser Hierarchie gelten Vererbungsprinzipien, zum Teil werden die Hierarchien in mehreren Dimensionen angelegt, sodass das monohierarchische Prinzip durchbrochen wird, oder es gibt verschiedene lokale, aber keine übergeordnete integrierende Taxonomie (vgl. Fridman Noy/Hafner:68). Der Aufbau wie auch die vorgesehenen Relationen zwischen den Konzepten/Termen richtet sich natürlich auch nach dem Zweck, den eine Ontologie erfüllen soll (vgl. Uschold/Gruninger:113). Eine Metaontologie legt fest, welche Elemente in der Ontologie enthalten sein können, über die dann Aussagen gemacht werden können (vgl. Köpcke 2002:330).

⁵ Vgl. dazu auch Weinstein/Birmingham, die eine entsprechende Skala für Vokabularstrukturen anführen, an deren informellem Ende bibliothekarische Klassifikationen rangieren (Weinstein/Birmingham:22).

⁶ Es gibt auch nicht taxonomische Ontologien, sie sind jedoch selten und oft auf sehr spezifische Bereiche zugeschnitten (vgl. Fridman Noy/Hafner:68)

In der Taxonomie sind Konzepte organisiert, die zum einen intern durch bestimmte Eigenschaften und Attribute strukturiert sein können und zum anderen in verschiedenen Relationen zueinander stehen. Die prominenteste ist die Teil-Ganzes-Beziehung, die weiterhin unterschiedlich fein unterteilt sein kann (vgl. Fridman Noy/Hafner:71), aber hier sind vielfältige und z.T. sehr spezifische Relationen möglich. Die Taxonomie umfasst zumeist Klassen, denen individuelle Vertreter als Instanzen zugeordnet sind. Darüber hinaus verfügen Ontologien über Axiome, in denen grundlegende Relationen, Inferenzregeln wie auch Einschränkungen formuliert sind. Je nach angewandeter Logik sind die Inferenzregeln und Beschränkungen aber auch in der Konzeptstruktur enthalten (vgl. Fridman Noy/Hafner:71f). In vielen Fällen gibt es Möglichkeiten, Modalität, Wahrscheinlichkeiten und Kontextbezogenheit auszudrücken. Die grundlegenden Logiken, in denen Ontologien ausgedrückt werden, sind Prädikatenlogik, Beschreibungslogik oder Frame-basierte Ansätze (vgl. auch Fensel:83ff).

Allerdings ist die praktische Ausgestaltung der einzelnen Ontologien wie auch die theoretischen Ansichten über Nützlichkeit und Notwendigkeit der verschiedenen einzelnen Eigenschaften und Relationstypen sehr verschieden; nicht zuletzt ist die verwendete Programmiersprache wie auch die zugrunde liegende Logik hierfür von grundlegender Bedeutung. Was jedoch im Einzelnen das zentrale Element von Ontologien ist, ist strittig: Für die einen sind es klar definierte IS-A-Relationen (vgl. Weinstein/Birmingham:22), für andere ist es die Ableitbarkeit und eindeutige Beschreibung der Konzepte v.a. durch ihre Positionierung in einem semantischen Netz (dies war der ursprüngliche Ansatz in WordNet, vgl. Miller u.a.:240), für noch andere ist es die eindeutige, formale Definition der Konzepte (vgl. Uschold/Gruninger:97) oder die Vielfalt der semantischen Beziehungen. Substantielle allgemeine Aussagen zu Aufbau und Struktur von Ontologien lassen sich daher nur schwer machen.

Dagegen wird allenthalben Standardisierung, Wiederverwendbarkeit und Übertragbarkeit von Ontologien als Ziel in der Ontologie-„Community“ genannt. Angesichts der Heterogenität von Ontologien – nicht nur bezüglich des formalen und strukturellen Aufbaus, auch hinsichtlich der verschiedenen Zielsetzungen und Allgemeingrade – ist es nicht erstaunlich, dass ein wichtiges Forschungsziel die Entwicklung von Standards und gemeinsamen Formalismen und Werkzeuge darstellt (vgl. Fridman Noy/Hafner:53). Diese sollen v.a. der Interoperabilität von Software, Datenbanken und auch den verschiedenen bereichsspezifischen Ontologien dienen (vgl. SUO). Entsprechend gilt nach wie vor der Ruf nach einer entwickelten Methodologie zum Erstellen von Ontologien sowie nach Standards, welche Kategorien in einer Ontologie prinzipiell enthalten sein sollten und welche grundlegenden Objektklassen (z.B. Dinge, Prozesse, Relationen) berücksichtigt sein sollten (vgl. Fridman Noy/Hafner:73). Nur so können die immer wieder angestrebten Zeile von Wiederverwendbarkeit und Integrierbarkeit verschiedener Ontologien verwirklicht werden. Das Bestreben, einen Rahmen zum Vergleichen verschiedener Ontologien abzustecken, wie es Fridman Noy/Hafner in ihrem Beitrag anstreben, ist als Schritt in diese Richtung zu verstehen (Fridman Noy/Hafner: 53f, 73). Mittlerweile gibt es eine Reihe von Standardisierungsbestrebungen, sodass abzuwarten bleibt, welche sich durchsetzen werden (vgl. Fensel:112ff). Auch eine „ANSI Ad Hoc Group on Ontology Standards“ existiert (<http://www-ksl.stanford.edu/onto-std/>). Die zugrunde liegende Idee ist die modulare Organisation von verschiedenen Ontologien sowohl verschiedener Bereiche als auch verschiedener Ebenen (allgemein, spezifisch, aufgabenbezogen etc.), um je nach Bedarf zusammengestellt werden zu können. In diesem

Zusammenhang sind die seit Beginn der Ontologieentwicklung geforderten und auch mittlerweile aufzufindenden Verzeichnisse („libraries“) von Ontologien zu nennen (vgl. Uschold/Gruninger:125; Ding:382; z.B. <http://www.daml.org/ontologies/> oder <http://www-ksl.stanford.edu/knowledge-sharing/ontologies/>), die einen Überblick über existierende Ontologien und ihre Wiederverwendungsmöglichkeiten ermöglichen sollen.

2.5 Beispiele

Exemplarisch sollen hier drei prominente (und große) Ontologien vorgestellt werden: Cyc, WordNet und UMLS. Dabei kann es in diesem Rahmen nur um ein grobe und kurze Charakterisierung dieser drei Ontologien gehen, eine genauere, detailliertere Darstellung würde den Rahmen der vorliegenden Arbeit bei weitem sprengen. Ziel der Darstellung ist es, einen Eindruck von bekannten, konkreten Ontologien zu vermitteln.

2.5.1 Cyc

Cyc ist ein groß angelegtes Projekt seit Mitte der 1980er-Jahre, in dem „Commonsense-Wissen“, das menschlichem (alltäglichem) Schließen zugrunde liegende Alltagswissen, entsprechende Heuristiken und Daumenregeln in einer großen Ontologie erfasst werden soll, wo es für die Weiterverwendung in KI-Anwendungen bereit gehalten werden soll (vgl. Cyc; Lenat/Guha:30; Fridman Noy/Hafner:57). Der weit reichende Kontext ist, eine „global ontology of human knowledge“ (Lenat/ Guha:30) zu erstellen und dieses für Computer verarbeitbar zu machen. Auf diese Weise soll intelligentes Arbeiten von Computern ermöglicht werden (vgl. Fensel:14f). Enthalten sein soll „the bulk of the factual and heuristic knowledge, much of it usually left unstated, that comprises ‚consensus reality‘: the things we assume everybody already knows“ (Lenat/Guha:28).

Die Sprache, in der das in Cyc enthaltene Wissen repräsentiert wird, ist CycL, eine Form von Prädikatenlogik, die um eine Typisierung von Relationen (sowohl von Funktionen als auch von Prädikaten) erweitert ist, die Reifizierung von Formeln und Prädikaten zulässt, eine Kontextmodellierung mit Hilfe von so genannten Mikrotheorien ermöglicht (vgl. Fensel:86) und zudem auch Modalität und Sicherheit in der Aussage zulässt (vgl. Fridman Noy/Hafner:72). Durch die Kontextbindung in Mikrotheorien und die kontextspezifische Zuweisung von Wahrheitswerten werden auch zunächst widersprüchliche Aussagen in der Wissensbasis ermöglicht (vgl. Fensel:86; Cyc). Die zentrale, hierarchiebildende Relation ist eine IS-A-Relation, in der die Mikrotheorien einzelne „Gitter“ („lattices“) bilden, mit denen verschiedene Perspektiven, Erschließungstiefen bzw. Granularitätsgrade etc. dargestellt werden können. Allerdings wird diese Anordnung mehrfach als verwirrend und unübersichtlich beschrieben (vgl. Fridman Noy/Hafner:57; Vickery:279).

Cyc stellt eine stark formalisierte Ontologie dar (vgl. Fensel:11), auf dem von Uschold/Gruninger vorgeschlagenen Kontinuum rangiert sie bei „sehr formal“.

Anwendung findet Cyc, auch wenn es primär zur Unterstützung und Weiterentwicklung von KI-Prozessen gedacht war (vgl. Chandrasekaran u.a.:25; Lenat/Guha:30), v.a. in Natural-Language-Anwendungen. Als zzt. erhältliche Anwendungen von Cyc werden u.a. die Integration und Abfrage von heterogenen Datenbanken, Retrieval von Nicht-Text-Materialien, Integration verschiedener Thesauri und Information Retrieval im WWW genannt (vgl. Cyc Applications; siehe auch Amann u.a.:223).

Die im Cyc-Projekt erarbeitete große, domänenunabhängige Ontologie ist allerdings nur im „upper level“ (der hier die oberen, allgemeinen 6000 Konzepte erfasst, vgl. Opencyc) öffentlich zugänglich, die spezifischeren Ausarbeitungen werden von der Firma Cycorp kommerziell vermarktet (vgl. Fensel:15).

2.5.2 WordNet

WordNet dagegen ist eine v.a. linguistisch motivierte Ontologie, die ebenfalls bereichsunabhängig ist und in der eine große Anzahl von Wörtern mit ihren Bedeutungen verzeichnet ist. Ziel der Ontologie ist, lexikalisches Wissen zu erfassen und maschinell verarbeitbar zu repräsentieren, um so dieses Bedeutungswissen für das Natural Language Processing nutzbar zu machen. Gedacht ist hier an Projekte im Information Retrieval, in maschineller Übersetzung und künstlicher Intelligenz (vgl. Fellbaum:2f).

Die zentrale Einheit in WordNet ist das Wort bzw. die mit Wörtern assoziierten kognitiven Konzepte; Wörter bzw. ihre Bedeutungen werden weder weiter analysiert noch werden größere kognitive Einheiten wie Frames angenommen, wohl aber Mehr-Wort-Begriffe, sofern sie atomare Konzepte repräsentieren (vgl. Fellbaum:4ff; Fridman Noy/Hafner:58, 70). Die Konzepte werden in so genannten „synsets“ („synonym sets“) repräsentiert, die die zentrale Einheit für die Organisation und Strukturierung von WordNet darstellen (vgl. Miller u.a.:237; Fellbaum:7; Fridman Noy/Hafner:57), und weniger die Wortform selbst, wie in Wörterbüchern üblich. Es geht nicht wie bei Cyc um die Darstellung von Welt- oder enzyklopädischem Wissen, Objekt des Interesses ist das mit zumeist einem Wort verbundene, atomare Konzept (vgl. Fellbaum:6). Grundannahme ist, dass Konzepte durch die Interaktion des Menschen mit seiner Umwelt entstehen und ggf. sogar angeboren und universal sein können. Da nur existierende Konzepte lexikalisiert werden, kann es vorkommen, dass in WordNet Konzepte aufgenommen werden, die im Englischen nicht lexikalisiert sind, evtl. aber in einer anderen Sprache, die also lexikalische Lücken des Englischen sind. Diese Grundannahme ist auch die Basis für die postulierte Sprachunabhängigkeit von Konzepten (vgl. Fellbaum:6, 8).

Konzepte werden durch die sie repräsentierenden Synsets in der lexikalischen Struktur positioniert, sie tragen zur Disambiguierung polysemer Wortformen bei, indem bestimmte Wortformen nur in bestimmten Kontexten synonym sind – entsprechend ist die Definition von Synonymie „abgeschwächt“ kontextrelativ formuliert (vgl. Miller u.a.:240f).

Obwohl ursprünglich nicht für notwendig erachtet (vgl. Miller u.a.:240) werden die Synsets durch Definitionen und Beispiele erläutert, ohne dass das damit vermittelte nicht-lexikalische, sondern enzyklopädische Wissen als Teil der lexikalischen Struktur aufgefasst werden soll (vgl. Fellbaum:6). WordNet unterteilt die von ihm repräsentierten Wortformen und Begriffe in die vier syntaktischen Kategorien Nomina, Verben, Adjektive und Adverbien (letzte wurden erst später aufgenommen)⁷.

Zwischen den Synsets existieren verschiedene semantische Relationen die durch Zeiger symbolisiert werden und reziprok sind. Sie werden – zumindest theoretisch – von rein lexikalischen Relationen wie Synonymie unterschieden (vgl. Miller u.a.:239, 240). Einige dieser Relationen wirken für je eine der o.g. Kategorien strukturgebend: Hyponymie für Nomina, was einer IS-A-Relation entspricht und zu einer hierarchischen Taxonomie führt, Meronymie für Verben, Antonymie für Adjektive (vgl. Miller u.a.:242f; Fridman Noy/Hafner:67).

⁷ Funktionswörter sind offensichtlich auch darin enthalten (vgl. Fensel:14).

WordNet ist wenig formalisiert, d.h. es bietet keine formalen Definitionen seiner Konzepte (wie z.B. Cyc), was zu Grenzen beim Einsatz in Rechnersystemen führt (vgl. Fensel:14). WordNet ist unter Lizenzierung frei erhältlich.

2.5.3 UMLS

Das Unified Medical Language System (UMLS) dagegen stellt ein bereichsspezifisches System dar, das ebenfalls seit Mitte der 1980er-Jahre von der US-amerikanischen National Library of Medicine (NLM) entwickelt wurde, um die Integration von und das Retrieval in verschiedenen biomedizinischen Informationsquellen zu vereinfachen (vgl. Fridman Noy/Hafner:61). Hauptanliegen ist, Probleme, die durch die Verwendung verschiedenartiger Terminologien und das Verstreut-Sein von relevanten Informationen auf verschiedene Datenbanken entstehen, zu überwinden (vgl. UMLS Fact Sheet).

Das UMLS basiert auf drei Komponenten: dem „Metathesaurus“, dem „UMLS Semantic Network“ und dem „Specialist Lexicon“.

Der Metathesaurus ist ein medizinischer Thesaurus, der Einträge aus verschiedenen kontrollierten Vokabularen und Klassifikationen, die für einschlägige Daten existieren, beinhaltet. Dabei will er die in den Original-Vokabularen enthaltenen strukturellen und semantischen Informationen übernehmen, zugleich neue Basisinformationen zu den Konzepten hinzufügen und Relationen zwischen den Konzepten aus verschiedenen Quellen erstellen (vgl. UMLS Fact Sheet Metathesaurus). Organisierendes Element sind die Konzepte, alternative Bezeichnungen für dasselbe Konzept sind verlinkt.

Der Metathesaurus wird ergänzt durch ein semantisches Netzwerk, das die Einträge des Metathesaurus semantischen Typen (z.B. „organism“ oder „biologic function“) zuweist und diese untereinander in Relation setzt. Neben den die einzige vorgesehene Hierarchie bildenden IS-A-Relationen, die auf der Basis-Dichotomie von „event“ und „entity“ aufbaut, gibt es hier verschiedene weitere, z.T. sehr Medizinspezifische Relationen (z.B. „treats“ oder „prevents“), die ihrerseits in einer Taxonomie zusammengefasst sind. Die Konzepte des UMLS sind atomar, sind also in sich nicht weiter strukturiert; Axiome, die Schlussregeln zugrunde liegen, sind ebenfalls nicht vorgesehen (vgl. Fridman Noy/Hafner:69ff; Köpcke 2001:Anhang 3; UMLS Fact Sheet Semantic Network).

Als weitere Ergänzung zu den beiden ersten Komponenten existiert ein „Specialist Lexicon“, das die lexikalische Basis für ein (englisches) Natural-Language-Processing-System darstellt, in dem sowohl allgemeinsprachliches wie auch biomedizinisches Vokabular enthalten ist. In diesem sind v.a. linguistische Hinweise verzeichnet, d.h. syntaktische, morphologische und auch orthografische Regeln und Varianten wie Wortbildungsregeln und Phrasen (vgl. UMLS Fact Sheet Specialist Lexicon). Damit soll die maschinelle Verarbeitung der Einträge im Metathesaurus, aber auch von medizinischen Texten erleichtert werden.

Angewendet wird das UMLS in den eigenen Anwendungen der National Library of Medicine wie dem öffentlichen Zugang zu Medline, PubMed, oder dem NLM Gateway. Es findet Einsatz im Information Retrieval über verschiedene Datenbanken, bei der Erstellung von Patientendaten, in der Indexing Initiative der NLM etc. (vgl. UMLS Applications).

Das UMLS ist frei erhältlich, die National Library of Medicine ist an einer Verbreitung selbst interessiert und erarbeitet regelmäßig Updates. Es ist jedoch eine Lizenzierung erforderlich.

2.6 Fazit

Ontologien stellen – ähnlich wie die klassischen semantischen Hilfsmittel im Information Retrieval: Thesaurus und Klassifikation – Begriffssysteme dar, die das (menschliche) Wissen über einen Ausschnitt der Welt repräsentieren sollen, und zwar in Form von Konzepten und Relationen zwischen diesen Konzepten. Diese Konzepte sind mehr oder weniger formal definiert, die zentrale, tragende Relation ist meistens eine hierarchische Subsumptionsbeziehung (IS A), es können jedoch zahlreiche andere, je nach Ontologie und Themengebiet differierende Relationen zwischen den Konzepten vorkommen. Zudem verfügen Ontologien über Schlussverfahren, seien sie über Axiome formuliert, seien sie aus den Definitionen der Relationstypen oder Konzepte ableitbar, die es ihnen ermöglichen, über die in den ontologischen Klassen verzeichneten Instanzen Aussagen (z.T. mit variierenden Wahrheitswerten) zu machen.

Zugleich stellen Ontologien „Wörterbücher“ dar, insofern sie Bezeichnungen der natürlichen (Fach)Sprache für die repräsentierten Konzepte und Relationen enthalten, auf diese Bezeichnungen auch für die Identifizierung, z.T. – in den weniger formalen Ontologien – sogar für die Definition der Konzepte zurückgreifen. Ontologien stellen für den jeweiligen Weltausschnitt Terminologie bereit.

Der erhoffte Nutzen von Ontologien ist weit gefächert: Neben der Bereitstellung von Weltwissen für problemlösende KI-Software und der Aufgabenpräzisierung und Konsistenzprüfung in der Systementwicklung finden sie Anwendung in der Wirtschaft – im E-Commerce, in firmeninterner Kommunikation als „corporate taxonomy“⁸ oder als Suchhilfe im Intranet –, sie stellen im Natural Language Processing Terminologie und Semantik und – hier von besonderem Interesse – versprechen im Information Retrieval intelligentere Suchmöglichkeiten und terminologische Anpassung und Integration sowohl im WWW, in Intranets oder über verschiedene Datenbanken hinweg.

Die entscheidenden Vorteile von formalen Ontologien gegenüber anderen Systemen werden zum einen in ihrer Verarbeitbarkeit durch Maschinen gesehen, die damit intelligentes Arbeiten simulieren können, d.h. also der Formalisierung und der Möglichkeit zu Inferenzen. Zum anderen werden sie in ihrer semantischen Reichhaltigkeit gesehen, worunter vielfältige Beziehungen zwischen den Konzepten sowie Kontextsensitivität in der Bedeutungsbeschreibung der Termini und die Widerspiegelung verschiedener Perspektiven verstanden wird.

Ontologien gibt es für sehr verschiedene Bereiche und für sehr verschiedene Zwecke, von der Darstellung menschlichen Alltagswissens (Cyc) über die Interpretation sprachlicher Ausdrücke (WordNet) bis hin zur Darstellung fachspezifischer Konzepte und Integration von Terminologie (UMLS). Es existieren viele, z.T. kleine und sehr spezifische Ontologien.⁹

⁸ Zu diesem Begriff vgl. Aitchison u.a.:7.

⁹ Einen Überblick über verschiedene Ontologie-Projekte gibt <http://www.cs.utexas.edu/users/mfkb/related.html>.

3 Ontologien in digitalen Bibliotheken

3.1 Digitale Bibliotheken

Eine digitale Bibliothek kann beschrieben werden als

„Einrichtung, die Texte, Bilder, Animationen, Ton- und Videoaufnahmen auf elektronischen Datenträgern verteilt (elektronische Bibliothek), die eine Vielzahl von Bibliotheksdiensten in einem ortsübergreifenden Verbund anbietet (virtuelle Bibliothek) und deren Dienste integriert sind, die einen effizienten Zugriff darauf über eine einheitliche Systemoberfläche gestattet und deren ‚Systemintelligenz‘ über die der Teile hinausgeht.“ (Endres/Fellner:4)

Dieses Zitat stellt eine von verschiedenen Beschreibungen dessen dar, was unter digitalen Bibliotheken zu verstehen ist; andere Beschreibungs- und Definitionsversuche setzen andere Schwerpunkte und Abgrenzungen zu benachbarten Begriffen (vgl. z.B. Rusch-Feja, Kap. 1).

Die hier interessanten Aspekte von digitalen Bibliotheken, die sie auch von konventionellen Bibliotheken unterscheiden, werden in dem oben angeführten Zitat deutlich: Zum einen hält eine digitale Bibliothek selbst elektronische Dokumente¹⁰ vor, die auch für andere Such- und Erschließungsstrategien zugänglich sind als konventionelle Medien (Volltextsuche, automatisches Indexieren). Zum anderen greift eine digitale Bibliothek in ihrem Dienstleistungsangebot nicht nur auf eigene Ressourcen zurück, sondern greift im Verbund mit anderen Organisationen (Bibliotheken, Verlagen, Informationszentren etc.) auch auf deren Angebote und Ressourcen zu, was bedeutet, dass sie für diese Zugriffe auch die verschieden aufgebauten und strukturierten Daten und Softwaresysteme koordinieren muss. Diese Dienste sollen zum Dritten integriert unter einer gemeinsamen Oberfläche angeboten werden, sodass die Distribution der Anfragen auf Datenquellen mit verschiedenen Anforderungen, Strukturen und Terminologien wie auch eine entsprechende Integration der jeweiligen Ergebnisse vonnöten ist.

Diese Aspekte sind zentral für digitale Bibliotheken, auch von anderen Autoren wie z.B. Rusch-Feja werden sie als zentrale Aufgaben beschrieben (vgl. Rusch-Feja, Kap. 5). Sowohl die Verwaltung (Erschließung und Recherche) von Volltexten als auch der Zugriff auf verschiedene Datenbanken und die Integration von Daten und Terminologien sind Bereiche, in denen Ontologien Anwendung finden können und existierende Verfahren deutlich verbessern sollen.

Um Möglichkeiten und Bereiche eines Einsatzes von Ontologien für Inhaltserschließung und Recherche in digitalen Bibliotheken besser einordnen zu können, sollen nun zunächst Anforderungen an das Information Retrieval in digitalen Bibliotheken angesprochen werden und kurz auf die mittlerweile als Standard geltende Auszeichnungssprache XML (eXtensible Markup Language) eingegangen werden.

¹⁰ Der Ausdruck (elektronisches) Dokument wird im Folgenden für alle in einer digitalen Bibliothek vorgehaltenen digitalen Objekte benutzt, unabhängig davon, ob es sich um Texte, Bilder, Animationen o.a. handelt.

3.1.1 Information Retrieval

Unter Information Retrieval (IR) wird

„die Repräsentation, Speicherung und Organisation von Informationen und der Zugriff zu Informationen“ (Salton/McGill:1)

verstanden.

Die bedeutendste Form des Information Retrieval in digitalen Bibliotheken ist das Textretrieval (vgl. Arms:160ff). Auch Bilder oder Audioaufnahmen werden häufig über Metadaten, die ihrerseits wieder eine Art von Text darstellen, erschlossen und recherchiert (Arms:188); reines Bild- oder Tonretrieval soll im Folgenden (ebenso wie das Retrieval von Fakten in besonderen Formaten wie z.B. Formeln) unberücksichtigt bleiben.

Konventionelles Retrieval von Literaturangaben in Literaturdatenbanken (oder auch Bibliothekskatalogen) erfolgt über Metadaten entweder formal nach Autor, Titel und ähnlichen Angaben sowie über semantische Hilfsmittel. Die semantische Suche kann über eine Stichwortsuche in Titel oder Abstract und, sofern ein Dokument digital erfasst ist, auch im Volltext erfolgen. Suchbegriffe können dort mit Hilfe von Thesauri gewichtet werden. Mit dem Einsatz moderner probabilistischer Retrievalmodelle gehen Rankingverfahren bei der Ergebnisbewertung einher. Des Weiteren erfolgt semantische Suche mit Hilfe von Thesauri und Klassifikationssystemen in den eigens erstellten deskriptiven semantischen Metadaten. Auch in digitalen Bibliotheken ist dies nicht grundsätzlich anders, auch hier wird nach Metadaten gesucht (vgl. Endres/Fellner:166), auch wenn die Volltextsuche natürlich ebenfalls eine Rolle spielt.

Besonderheiten ergeben sich v.a. dadurch, dass digitale Bibliotheken wie erwähnt bestrebt sind, die Abfrage (wie auch die Ergebnisausgabe) von verschiedenen Quellen gleichzeitig und einheitlich zu ermöglichen und dabei die vorhandenen heterogenen Datenquellen integrieren müssen. Digitale Bibliotheken bieten in der Regel ein „abgestimmtes Angebot von Suchhilfen“ an, das sich über Kataloge der eigenen (auch nicht digitalen) Bestände, über andere Referenz-Datenbanken, Suchmaschinen für das eigene Volltextangebot wie auch Suchmöglichkeiten von lizenzierten Volltexten anderer Provenienz, ggf. auch Internetsuchmaschinen und -kataloge erstreckt (vgl. Endres/Fellner:182)

Für eine einheitliche und ungebrochene Suchprozedur wie auch Ergebnisausgabe ist die Transformation der Datenbankabfragen erforderlich, die sich auf Daten-Modell, Abfrage-Sprache wie auf semantische Deskriptoren und Klassifikationen erstreckt. Dazu kommt das Mischen und Sortieren der Suchergebnisse aus den verschiedenen Quellen, die von Dublettenelimination bis hin zur ausgesprochen schwierigen Integration verschiedener Ranking-Strategien reichen kann. Auch multilinguales Retrieval ist hier zu erwähnen.

Zur Verknüpfung der vielen und heterogenen Quellen für die Antworten von Suchanfragen werden Mediatoren eingesetzt. Dies sind Softwarekomponenten, die die Abfragetransformationen wie auch die Ergebnisintegration bewerkstelligen sollen. Sie erfordern jedoch immer wieder intelligente Interpretationsleistungen und sind daher nur begrenzt einsetzbar. Z.T. wird versucht, komplexe Suchanfragen mit Hilfe von lernfähigen Softwareagenten zu bearbeiten, die einzelne Aufgaben in der komplexen

Suche übernehmen sollen (vgl. Endres/Fellner:401). Rusch-Feja sieht für die entsprechende Systemarchitektur Ontologien als zentral an (vgl. Rusch-Feja, Kap. 5.1)

Zentrale Probleme der Recherche in semantischen Metadaten sind, dass die klassischen semantischen Hilfsmittel Thesaurus und Klassifikation eine genaue Kenntnis erfordern, um sie effektiv einsetzen zu können, verschiedene, auch themengleiche Thesauri aber eine unterschiedliche Terminologie haben, verschiedene Datenbanken zudem noch unterschiedliche Abfragesprachen.

Ein neuerer Ansatz ist hier die „Konzeptsuche“, die allerdings ebenfalls Verfahren zur Identifizierung der Konzepte voraussetzt (vgl. Endres/ Fellner:401)

Durch elektronisch vorliegende Thesauri und Klassifikationen werden die Suchmöglichkeiten durch Navigieren und Browsen erweitert (vgl. Arms:187; Endres/Fellner:170ff; Köpcke 2002:325). Thesauri ermöglichen darüber hinaus, Synonyme zu erkennen, und können alternative und spezifischere Begriffe zum eigentlichen Suchbegriff bereitstellen (vgl. Arms:203; Köpcke 2002:325). Probleme, Wörter identifizieren und von anderen, gleich geschriebenen unterscheiden zu können, werden durch elektronische Wörterbücher und so genannte „stemmer“ gelöst, die – anders als einfache Trunkierung –, Wörter auf ihre Wortstämme zurückführen und dabei Homonyme erkennen können (vgl. Arms:203). Damit werden Thesauri und Klassifikationen auch zunehmend für Endnutzer von Bibliotheken benutzbar und finden Eingang in die Internet- und Volltextsuche.

Für die Suche in Volltexten ist es wünschenswert – wenn auch zzt. nicht realistisch –, in den Texten navigieren zu können, um so die relevanten Textpassagen schneller zu finden (vgl. Endres/Fellner:183).

Die Informationen, die von Nutzern in digitalen Bibliotheken gesucht werden, sind dabei ausgesprochen heterogen. Hier kann allgemein nach Überblicksinformation zu einem Thema wie auch eine ganz bestimmter Text oder eine Textpassage, verschiedenartige Fakten oder die gesamte relevante oder auch sehr spezifische Literatur in einem Gebiet gesucht werden (vgl. Arms:187f). Zudem suchen die Benutzer von digitalen Bibliotheken zunehmend ihre Daten selbst – bedingt durch die verbreitete Nutzung von elektronischen Medien, insbesondere dem WWW und den dort üblichen und notwendigen Suchmaschinen einerseits, andererseits durch die immer benutzerfreundlicheren Recherchemöglichkeiten zum Beispiel durch digitale Bibliotheken. Die Aufgabe von Bibliothekaren liegt dementsprechend zunehmend in der kompetenten Beratung und Hilfestellung der Nutzer bei der Durchführung eigener Recherchen (vgl. Endres/Fellner:186).

3.1.2 XML

In digitalen Bibliotheken spielt die Auszeichnungssprache XML für die Speicherung und Übertragung von Texten und Metadaten eine zentrale Rolle. XML stellt eine vereinfachte Entwicklung von SGML (Standard Generalized Markup Language) dar, die sich durch ihre Einfachheit und durch den Verzicht auf besonders komplexe Konstrukte als Standard erfolgreich etablieren konnte. Die Flexibilität und Ausdruckstärke von SGML bleiben dabei zumindest in Teilen erhalten.

XML ermöglicht über seine „tags“ eine strukturelle Auszeichnung und Beschreibung der verschiedenen Textelemente, die unabhängig ist vom Layout des Textes. Das Layout wird seinerseits unabhängig vom eigentlichen Textkörper in einer eigenen Auszeichnungssprache (XSL) definiert. Damit kann einem Text oder Textteil je nach Bedarf ein verschiedenes Layout zugewiesen werden, z.B. je nach Ausgabe-medium (Bildschirm/Drucker) oder nach persönlichen Profilen (vgl. Fensel:76f).

XML ermöglicht darüber hinaus eine Strukturierung des jeweiligen Textes und eine individuelle oder (gruppen-/bereichsspezifische) Definition der Tags in DTDs (Document Type Definitions). Damit ist eine ausführliche strukturelle wie semantische Auszeichnung der Texte in verschiedenen Detaillierungsgraden möglich, die den jeweiligen Zielgruppen- oder Autoren- und Herausgeberbedürfnissen entspricht, für diese dann aber einheitlich definiert ist und dadurch Interoperabilität erlaubt (vgl. Arms:173ff).

XML ermöglicht also eine problemlose Integration von Metadaten. Als Standard für den Austausch und die Integration von verschiedenen Metadaten-Schemata (wie dem Dublin Core) ist RDF (Resource Description Framework) eine Möglichkeit, das in der Regel in XML ausgedrückt wird (vgl. Arms:198f, Endres/Fellner:300).

3.2 Einsatz von Ontologien

Formale Ontologien, die in digitalen Bibliotheken wie auch im WWW Einsatz finden, bedienen sich zum Teil der Möglichkeiten, die XML bietet; es existieren verschiedene XML- oder auch HTML-kompatible Repräsentationssprachen für Ontologien¹¹ (vgl. Fensel:102ff; Köpcke 2002:335; Köpcke 2001:48ff), Ontologien setzen zum Teil an RDF oder direkt an den Metadaten an (vgl. Fensel:97ff). Wie schon erwähnt, herrscht hier allerdings eine große Heterogenität, und auch die in digitalen Bibliotheken eingesetzten Ontologien basieren auf verschiedenen Grundlagen und Logiken.

In Folgenden sollen einige Projekte vorgestellt werden, die für verschiedene Aspekte versuchen, Ontologien in digitalen Bibliotheken einzusetzen.

In digitalen Bibliotheken werden Ontologien zum einen bei der Recherche eingesetzt: Zur Integration verschiedener Terminologien und Thesauri bei der Abfrage verschiedener Datenbanken innerhalb eines fachlichen Bereiches (z.B. UMLS), bei der Literatur- und Informationsrecherche, wobei z.T. andere, nicht unbedingt thematisch orientierte Herangehensweisen und Zusammenhänge dargestellt werden, und bei der Unterstützung von wissenschaftlichen Diskursen (vgl. Weinstein/ Birmingham; Welty/Jenkins; Buckingham Shum u.a.). Ziel scheint in jedem Falle zu sein, für die digitale Bibliothek einen „Mehrwert“ zu schaffen, der das Rechercheergebnis im Sinne der von Endres/Fellner genannten „Systemintelligenz“ mehr als die Summe seiner Teile sein lässt. Durch das Zusammenbringen verschiedener Informationen sollen Synergieeffekte erzeugt werden.

Auch bei der Inhaltsschließung werden Ontologien eingesetzt, z.B. beim Abstracting (Endres-Niggemeyer), und es wird auch bei der Erstellung von Metadaten, die als Konzepte in einer Ontologie recherchierbar sein sollen, auf eben diese Ontologie zurückgegriffen. Die Generierung von entsprechenden Metadaten erfolgt mit Hilfe der Ontologie z.T. (halb)automatisch (vgl. Weinstein/Birmingham).

Ontologien aber werden noch weit versteckter hinter den Kulissen eingesetzt, z.B. in Mediationssystemen (vgl. Amman u.a.) oder zur Koordination von Agentensystemen (vgl. Weinstein/Birmingham). Auf diese Ansätze kann aber im vorliegenden Zusammenhang nicht näher eingegangen werden.

Der konkrete Einsatz von Ontologien in digitalen Bibliotheken erfolgt meistens im Zusammenhang mit einzelnen Projekten, in denen die Möglichkeiten von KI-Anwendungen getestet werden sollen; nur selten ist die Nutzung der Ontologien etabliert (eine Ausnahme stellt das UMLS dar). Einige dieser Projekte sollen nun exem-

¹¹ Beispiele sind hier SHOE (Simple HTML Ontology Extensions) oder OIL (Ontology Inference Layer).

plarisches skizziert werden, um Einsatzmöglichkeiten von Ontologien in digitalen Bibliotheken aufzuzeigen.

3.2.1 Recherche

Auf den Aufbau des Unified Medical Language System (UMLS) bin ich in Kapitel 2 bereits kurz eingegangen. Hier soll nur noch einmal darauf hingewiesen werden, dass das Hauptanwendungsgebiet von UMLS auf der Integration verschiedener medizinischer Thesauri v.a. für Datenbankabfragen liegt. Die semantischen Relationen und Kategorien, die vom semantischen Netzwerk bereitgestellt werden, ermöglichen, nach Suchbegriffen zu suchen oder durch die kategoriale Hierarchie zu blättern (vgl. Köpcke 2002:333). Anwendung findet das UMLS neben Datenbankrecherchen (z.B. Medline) u.a. in der Suche in verschiedenen, heterogenen Datenquellen an der National Library of Medicine (Referenz-Datenbanken, Volltexte, Organisationsverzeichnisse etc.) und in der WWW-Volltextsuche. Darüber hinaus wird das UMLS noch in Projekten (der NLM) zum automatischen und halbautomatischen Indexieren eingesetzt (siehe Kap. 3.2.2; vgl. UMLS Applications).

Ein anderes Beispiel für die Anwendung von Ontologien im Bereich der Recherche ist ihre Nutzung für die thematische Klassifikation von Informationen in digitalen Bibliotheken. In einem Projekt am Vassar College (Staat New York) wurde im Zusammenhang mit einem Digitalisierungsprojekt historischer Quellen zum amerikanischen Bürgerkrieg versucht, formale Ontologien für eine Erweiterung der über der Klassifikation suchbaren Objekte zu nutzen (vgl. Welty/Jenkins). Ziel ist die Bereicherung der in einer Suchanfrage abrufbaren Informationen, suchbar sollten nicht nur die Literatur zu einem Thema¹² sein, sondern auch Hintergrundwissen wie Personen, Orte, Ereignisse, Organisationen etc., die für dieses Thema relevant sind (vgl. Welty/Jenkins:158f). Die Ontologie beruht auf Beschreibungslogik, die Instanzen der Hauptklassen der Ontologie, Ereignis, Objekt, Dokument (daneben gibt es noch Dokument-Modalität und Thema), können ihrerseits wieder nach dem Thema klassifiziert sein, sodass sie über eine allgemeine Themensuche recherchierbar sind (vgl. Welty/Jenkins:159). Zwischen den in den Klassen (außer der Klasse „Thema“) verzeichneten Individuen existieren vielfältige Beziehungen.

Ziel der Ontologie ist, im Volltext elektronisch erfasste und tief gehend intellektuell ausgezeichnete historische Quellen v.a. über Metadaten zugänglich und unter verschiedenen Hinsichten durchsuchbar zu machen. Es sollen die Möglichkeiten von Recherche und Browsing verbessert werden, insbesondere, wenn Nutzer nur vage Vorstellungen davon haben, was sie suchen, oder sich nur bruchstückhaft an ein einmal gesehenes Objekt erinnern. Die historischen Quellen sollen für vielerlei Forschungszwecke und den damit verbundenen, sehr unterschiedlichen Fragestellungen zugänglich gemacht werden.

Mit Hilfe der in der Ontologie enthaltenen Relationstypen und Inferenzregeln können Antworten zu Fragen generiert werden, ohne dass diese Elemente vorher explizit als eine Antwort zu eben dieser Frage definiert wurden (vgl. Welty/Jenkins:159f). Das darin enthaltene Regelwissen ist sonst oft nur implizit bei den Personen, die eine

¹² „Subject“, was von Welty/Jenkins explizit von „topic“ unterschieden wird (vgl. Welty/Jenkins: 161, 176f).

Klassifikation von Texten intellektuell vornehmen, vorhanden und soll durch die vorliegende Ontologie explizit gemacht werden (vgl. Welty/Jenkins:165).¹³

Die Ontologie wurde in Auseinandersetzung mit bibliothekarischen Klassifikationssystemen entwickelt, an denen v.a. eine primäre Ausrichtung an der Platzierung von Büchern in Regalen kritisiert wird. Mit der Ontologie sollen Informationen über verschiedenen Arten von Gegenständen (neben Dokumenten auch Personen, Organisationen, Ereignisse usw.), die in den Volltexten enthalten sind, thematisch klassifiziert und auffindbar gemacht werden (vgl. Welty/Jenkins:163ff). Das hat Implikationen für die Struktur und Repräsentation von „Thema“ (bzw. „subject“), da die darunter klassifizierten Gegenstände in sehr verschiedenen Relationen zu eben diesem Thema stehen – und dies ist in der Regel nicht die in objektorientierter Darstellung übliche Instance-Of-Relation (vgl. Welty/Jenkins:168ff).

Ein weiteres Beispiel, bei dem Ontologien eingesetzt werden, um ein attraktives Dienstleistungsangebot in digitalen Bibliotheken umzusetzen, ist „ScholOnto“, ein Digital-Library-Server-Projekt, mit dem wissenschaftlicher Diskurs anhand eines semantischen Netzwerkes ermöglicht werden soll (vgl. Buckingham Shum u.a.). Die Ontologie bezieht sich auf Formen wissenschaftlicher Argumentation und Auseinandersetzung, mit denen sich Wissenschaftler gegenseitig auf ihre publizierten Texte beziehen. Damit soll versucht werden, den Rückgriff auf disziplinspezifische Fachontologien zu vermeiden, bei denen die Gefahr fehlender Aktualität wie auch der Festschreibung bestimmter Perspektiven gesehen wird. Die fachlichen Bereiche werden anhand der Aussagen („claims“) der Autoren über ihre Arbeit dargestellt. Die daraus resultierende Ontologie umfasst dann Konzepte wie „Analysis“, „Approach“, „Idea“, „Methodology“, die enthaltenen Relationen heißen z.B. „confirms“, „refutes“, „uses“, „modifies“ (vgl. Buckingham Shum u.a.:240). Die Konzepte und Relationen sind in den „claims“ der Autoren organisiert, die den Angelpunkt der Ontologie ausmachen. Längerfristig soll auf diese Weise ein komplexes Netzwerk wissenschaftlicher „Schulen“ entstehen, in dem über Inferenzregeln Positionen und Perspektiven der zu der Ontologie beitragenden Wissenschaftler strukturiert und zusammengefasst werden (vgl. Buckingham Shum u.a.:241f).

Ziel ist es, mit der Ontologie eine Infrastruktur zu schaffen für explizite, zusammenfassende Beschreibungen von wissenschaftlichen Texten, die ein gezieltes Auffinden von inhaltlich zusammengehörigen, auf eben diese Weise der wissenschaftlichen Auseinandersetzung miteinander verbundenen wissenschaftlichen Dokumenten ermöglichen sollen. Es soll auch ein neues Medium wissenschaftlicher Diskussion entstehen, in dem sich Wissenschaftler der Ontologie bedienen, um Texte zu verfassen und miteinander in Disput zu treten (vgl. Buckingham Shum u.a.:240). Die Inhaltserschließung und Dokumentbeschreibung in diesem Projekt soll durch die sich beteiligenden Wissenschaftler selbst geschehen, die sich im wissenschaftlichen Disput positionieren und ihre Arbeit zu anderen in Beziehung setzen.

Als letztes Beispiel für die Nutzung von Ontologien im Bereich der Recherche in digitalen Bibliotheken soll die University of Michigan Digital Library (UMDL) dienen (vgl. Weinstein/Birmingham). Die hier verwendete Ontologie will Beziehungen zwischen einzelnen im Katalog verzeichneten Werken darstellen. Sie bezieht sich im

¹³ Die Antwort auf solche Fragen wird dabei allerdings nur aus den im vorliegenden Textkorpus enthaltenen Informationen generiert, bezieht sich also nicht auf das in einem Gebiet vorhandene Wissen, wie es sonst von Ontologien oft beansprucht wird.

Wesentlichen auf eine entstehungsgeschichtliche „Werk-Hierarchie“, von der Idee bis zum Exemplar der gedruckten Veröffentlichung, die auf den Vorschlägen der IFLA zu den „functional Requirements for bibliographic records“ beruht und der die Attribute der Ontologie zugeordnet sind. Diese Attribute beschreiben verschiedene bibliografisch relevante Einheiten wie Autor, Titel, Genre, Format, Exemplar-Charakteristika (z.B. die Adresse in der digitalen Bibliothek) (vgl. Weinstein/Birmingham: 23f). Die kompletten Metadaten für ein konkretes Exemplar beinhalten einen Instance-Of-Eintrag auf jeder Hierarchie-Ebene. Auf diesen hierarchischen Ebenen sind auch die verschiedenen Relationen zwischen Werken (Übersetzung, Reihe etc.) angesiedelt (vgl. Weinstein/Birmingham:25). Sowohl die Konzepte als auch die Relationen werden hier aus (formalen) MARC-Datensätzen generiert, es sollen in MARC implizite Relationen, die sich z.B. aus gemeinsamen Metadaten oder natürlichsprachigen Kommentaren ableiten lassen, explizit gemacht werden (vgl. Weinstein/Birmingham:21). Ziel der Ontologie ist, diese Beziehungen zwischen einzelnen Werken, Ausgaben u.Ä. zu verdeutlichen und die Werke anhand dieser Relationen recherchierbar zu machen. Dabei wird die klassische Trennung von Zugangs- und Inhaltsdaten durchbrochen (Weinstein/Birmingham:26). Gesucht werden kann in dieser Ontologie auf verschiedenen Ebenen, ein Werk, z.B. die 5. Sinfonie von Beethoven¹⁴, kann allgemein in irgendeiner Interpretation, von einem bestimmten Interpreten/Orchester, in einer bestimmten Aufnahmeart, in einer ganz bestimmten Aufnahme und gemeinsam mit Noten oder entsprechender Literatur gesucht werden. Die Ontologie soll hier dank besserer Querverweise und Relationierung eine bessere und präzisere Suche, die Präzisierung vager Anfragen, die Integration verschiedener Kataloge und (bei Musikalien von besonderer Bedeutung) eine Vorkalkulation evtl. notwendiger Lizenzen ermöglichen (vgl. Weinstein/Birmingham:26). Auch diese Ontologie wird in einem beschreibungslogischen System dargestellt (vgl. Weinstein/Birmingham:22).

3.2.2 Erschließung

In der von mir herangezogenen Literatur werden Ontologien eher zurückhaltend in der Inhaltserschließung eingesetzt, es werden vornehmlich Studien und Projekte vorgestellt, die sich auf die Optimierung und/oder Erweiterung der Recherchemöglichkeiten konzentrieren (vgl. Kapitel 3.2.1). Eine Reihe von Projekten widmet sich auch der Ausgestaltung der Systemarchitektur digitaler Bibliotheken auf die hier aber nicht näher eingegangen werden soll (vgl. aber Kapitel 3.2.3).

Die Zuordnung ontologischer Konzepte u.Ä. zu Dokumenten oder Dokumentteilen erfolgt in den Beispielen häufig intellektuell. Ontologien werden bestenfalls eingesetzt, um Beziehungen zwischen Konzepten zu generieren, die Zuweisung der Konzepte selbst erfolgt, wenn automatisiert, über andere Verfahren. Im Beispiel der UMDL ist es ein Mapping-Verfahren (s.u.), ein anderes Beispiel liefern Stuckenschmidt/van Harmelen, die ein syntaktisches automatisches Klassifikationsverfahren verwenden, das um einen Ontologie-basierten „Reasoner“ ergänzt wird, der (Unterklassen-)Relationen zwischen Individualkonzepten einfügt (vgl. Stuckenschmidt/van Harmelen:167). Bei Stuckenschmidt/van Harmelen wird die Ontologie noch zur Konsistenzprüfung herangezogen.

Das UMLS allerdings findet auch in einem Indexierungsprojekt der National Library of Medicine Einsatz, in dem medizinische Texte verschiedener Provenienz (Web-Seiten, Aufsätze, Patientenberichte usw.) indexiert werden. Den Dokumenten

¹⁴ Das verwendete Beispiel-Korpus bezieht sich auf musikalische und musikwissenschaftliche Werke.

werden letztlich Terme der Medical Subject Headings (MeSH) zugeordnet, also nicht die Konzepte der Ontologie selbst. Die Aufgabe der UMLS-Ontologie besteht u.a. darin, verschiedene Terminologien zusammenzuführen, Synonyme und andere eng assoziierte Bezeichnungen zu erkennen und diese eben auf MeSH-Terme zurückzuführen (vgl. NLM Indexing Initiative).

Das bereits genannte UMDL-Projekt (vgl. Kapitel 3.2.1) versucht, mit Hilfe der Ontologie ontologische Metadaten aus vorliegenden MARC-Datensätzen zu generieren. Dazu werden die MARC-Daten in einen in einer Auszeichnungssprache strukturierten Text konvertiert, der Konzepte der UMDL-Ontologie enthält, die durch ein Mapping-Verfahren den MARC-Feldern zugeordnet sind. Dieser Text wird wiederum in die Repräsentationssprache der UMDL-Ontologie konvertiert, sodass die MARC-Felder und -Werte nun als ontologische Konzepte erscheinen. Dieser Pool an ontologischen Konzepten wird schließlich einer Reihe von Schluss- und Inferenzregeln, die in der Ontologie enthalten sind, unterzogen, um die für die Recherche letztlich gewünschten Relationen zu erhalten. Dabei wird letztlich zwischen wahrscheinlichen und sicheren inferierten Relationen unterschieden (vgl. Weinstein/Birmingham:27).

Mit der Bereicherung der formalen MARC-Metadaten um ontologische Konzepte und Relationen, die sich auf die Produktion von Werken auf den verschiedenen Allgemeinheiten- und Produktionsstufen beziehen und Bezüge zwischen einzelnen Werken herstellen, ist ein Schritt getan in Richtung der Inhaltserschließung von bibliothekarischen Dokumenten mit Hilfe einer Ontologie, bei der ontologische Schlussregeln und Inferenzen zur Erstellung von inhaltlichen Beschreibungen der Dokumente herangezogen werden.

Auch für Verfahren des automatischen Zusammenfassens werden Ontologien herangezogen. Ein Beispiel dazu kommt ebenfalls aus dem medizinischen Kontext: Die automatische Zusammenfassung zur Knochenmarkstransplantation aus dem WWW (vgl. Endres-Niggemeyer). Hier wird eine kleine, fachspezifische Ontologie benutzt, um in einem Agentensystem Fakten und Sprachverarbeitungsregeln zur Verfügung zu stellen; damit dient sie der Relevanzbewertung von Textpassagen. Sie stellt neben sprachlichen Synonymen hierarchische Begriffsbeziehungen und Kontextualisierungen zur Verfügung und ist v.a. im Zusammenhang mit Verfahren des Natural Language Understanding zu sehen. Zur Ankopplung an größere Systeme enthält die hier benutzte Ontologie Verweise auf die allgemeine Ontologie WordNet wie auch auf den zentralen Fachthesaurus MeSH (vgl. Endres-Niggemeyer:8).

3.2.3 Einbettung in die Systemarchitektur

Nur kurz erwähnt werden soll hier, dass Ontologien durchaus in komplexen Systemarchitekturen von digitalen Bibliotheken Platz und Anwendung finden, wo sie auf verschiedenen Ebenen v.a. die Interoperabilität verschiedener (Datenbank-)Systeme gewährleisten sollen.

Genannt sei hier als Beispiel noch einmal die University of Michigan Digital Library (UMDL), in der in einem Agentensystem die (Dienst-) Leistungen der Agenten mit Hilfe einer Ontologie klassifiziert werden, so dass das Dienstleistungsangebot der digitalen Bibliothek mit Hilfe dieser Ontologie während der Ausführung gesteuert wird (vgl. Weinstein/ Birmingham:29ff).

Ein anderes Beispiel ist der Ansatz, Ontologien in Kombination mit Thesauri zur Erstellung von RDF-Schemata einzusetzen, die in Mediatoren bei der Abfrage von Metadaten verschiedener XML-Quellen anwendungs- oder domänenspezifisch eingesetzt werden sollen (vgl. Amann u.a.).

3.3 Fazit

Betrachtet man die in digitalen Bibliotheken an Recherche und Inhaltserschließung gestellten Anforderungen, ist es nahe liegend, die Einsatzmöglichkeiten von Ontologien in diesem Bereich zumindest zu prüfen.

Hinsichtlich der Recherche sind Ontologien sicherlich als Mittel einer „Konzeptsuche“ anzusehen. Sie können Einsatz finden zum einen in der Volltextsuche, v.a. der von digitalen Bibliotheken selbst vorgehaltenen Textdokumente, aber auch für die Suche in (ontologischen) Metadaten von entsprechend ausgezeichneten Dokumenten und der WWW-Suche.¹⁵ Zum anderen können Ontologien bei der Integration von Abfragen verschiedener Datenbanken genutzt werden, nicht nur hinsichtlich von verschiedenen Datenstrukturen, sondern v.a. im Hinblick auf die Integration verschiedener Terminologien und Thesauri. Als Mittel zur Erhöhung von Interoperabilität können Ontologien einerseits an der Oberfläche für die Nutzer offensichtlich erscheinen oder aber „versteckt“ in die Systemarchitektur der digitalen Bibliothek eingebettet sein.

Ontologien können auch eingesetzt werden, um Nutzer- und Suchbedürfnisse zu befriedigen, die mit neueren Entwicklungen wie digitalen Bibliotheken auftreten: Neben Möglichkeiten des Navigierens und Browsens, die Ontologien bieten können, ist hier die einheitliche Suche nach verschiedenen Arten von Informationen (Daten, Fakten, Literatur etc.) zu nennen, aber auch die Versuche, neuere, nicht ausschließlich am Sachgebiet orientierte Recherchezugänge zu bieten.

In Bezug auf die Inhaltserschließung können Ontologien in Kombination mit anderen Werkzeugen des Natural Language Understanding für automatische Klassifikation und Abstracting eingesetzt werden, v.a. aber auch für (halb-)automatisches Indexieren ontologischer Konzepte. Ebenso können sie genutzt werden, aus vorhandenen Daten ontologische Relationen zu generieren.

In den von mir herangezogenen Beispielen liegt der Schwerpunkt des Einsatzes von Ontologien auf der Recherche. Neben einem Einsatz als „Wörterbuch“ zur Terminologie-Integration (UMLS) und der systeminternen Verwendung bei der Schaffung von Interoperabilität (Amann u.a.) oder zur Steuerung von Agentensystemen (UMDL) steht hier die Eröffnung neuartiger Suchstrategien im Vordergrund, die auf bisher nicht oder kaum verzeichnete (ScholOnto) oder implizite (UMDL) Relationen zwischen Dokumenten zurückgreifen. Auch der Versuch, verschiedene Arten von Informationen einheitlich in einem Volltextkorpus recherchierbar machen zu wollen (Welly/Jenkins), kann im Zusammenhang mit der Befriedigung neuer Nutzerbedürfnisse gesehen werden.

Nicht alle Möglichkeiten des Einsatzes von Ontologien sind mit den von mir angeführten Beispielen abgedeckt, z.B. sind Möglichkeiten der Volltextsuche nur gestreift. Aber es fällt auf, dass einige der Beispiele auf ausschließlich intellektuell erschlossene Dokumente zurückgreifen, obwohl der Einsatz von Ontologien auch in der Inhaltserschließung nahe liegend wäre. Hier kann man vermuten, dass in diesem Bereich entweder kein Bedarf gesehen wird – was angesichts der Kosten von Inhaltserschließung unwahrscheinlich ist –, oder aber ein tatsächlicher Einsatz von Ontologien hier vor großen Schwierigkeiten steht. Sowohl intelligente Volltextsuche als auch Inhaltserschließung setzen Verfahren der Spracherkennung voraus, wie es zum Beispiel von Endres-Niggemeyer oder im UMLS auch eingesetzt wird, sodass auch nach hier Gründen gesucht werden kann.

¹⁵ Auf die Ansätze zu einem „semantic web“ (<http://www.w3.org/2001/sw/>) kann ich im Rahmen dieser Arbeit nicht eingehen, sie sind aber sicherlich für digitale Bibliotheken interessant.

Im folgenden Kapitel soll nun überlegt werden, inwieweit Ontologien den herkömmlichen Mitteln der Inhaltserschließung und Recherche, Thesaurus und Klassifikation, in diesen Bereichen überlegen sind.

4 Thesauri und Ontologien – alter Wein in neuen Schläuchen?

In diesem Kapitel sollen gefragt werden, inwieweit formale Ontologien für Recherche und inhaltliche Erschließung in digitalen Bibliotheken tatsächlich von Nutzen sind. Dabei kann es sich hier nur um eine Skizzierung von Ansatzpunkten handeln, anhand derer eine solche Überlegung weiter geführt werden könnte; genauere Urteile und Ergebnisse erforderten eine sehr viel genauere Analyse der jeweiligen Ontologien wie auch ihres konkreten Einsatzes unter konkreten Bedingungen, als sie in dieser Arbeit geleistet werden kann.

Insbesondere in Bezug auf das Information Retrieval, im Hinblick auf inhaltliche Erschließung und semantische Suchhilfen, sind Ontologien v.a. mit Klassifikationen oder Thesauri zu vergleichen.

Klassifikation bedeutet allgemein

*„eine Gruppierung oder Einteilung des gesamten Wissens, der Wissenschaft und ihrer Disziplinen nach einheitlichen methodischen Prinzipien.“
(Manecke:141)*

In einer Klassifikation werden Begriffe in ein „System der wechselseitigen Über- und Unterordnung eingebettet“ (Fugmann:43). Handelt es sich dabei um eine Abstraktionshierarchie, so ist die tragende hierarchische Relation eine IS-A-Beziehung (vgl. Fugmann:44). Nichts anderes tut eine Ontologie, die ebenfalls Begriffe systematisiert und anordnet, und zwar im Allgemeinen hierarchisch in einer IS-A-Beziehung.¹⁶ Entsprechend werden Ontologien und Klassifikationen auch verglichen oder sogar gleichgesetzt (vgl. Soergel 1999; Vickery:284; ebenfalls Welty/Jenkins, die ihre Ontologie in Auseinandersetzung mit Klassifikationsprinzipien entwickeln).

Eine Klassifikation versucht, tatsächlich das Wissen eines Gebietes oder auch insgesamt „der Welt“ systematisch zu repräsentieren, sie benutzt zur Darstellung oft eine künstliche Notationssprache, mit der die Klassenbegriffe bezeichnet und in Relation gesetzt werden und die durch natürlichsprachige Erläuterungen ergänzt wird. Alternativ kann auch eine bestimmte natürlichsprachige Bezeichnung für die Begriffe gewählt werden. Bei Klassifikationen geht es weniger darum, einen konsensualen Sprachgebrauch zu reflektieren, wie es für Thesauri gilt. Der Inhalt einer Klassifikation, die zur Inhaltserschließung benutzt wird, richtet sich auch nicht nach den Inhalten der zu dokumentierenden Dokumente – ausschlaggebend ist v.a. das in einem bestimmten Bereich vorhandene Wissen (vgl. Manecke:146). Dies ist, was auch Ontologien für sich beanspruchen: das in einem Bereich (oder auch der ganzen Welt) vorhandene Wissen zu erfassen und systematisch zu reflektieren.

Trotzdem scheint es mir sinnvoller zu sein, Ontologien mit Thesauri als mit Klassifikationen zu vergleichen. Ein Vergleich von Ontologien mit Thesauri ist zum einen nahe liegend, weil für die meisten Ontologien auch die Erfassung und die Integration der in einem Gebiet existierenden, mehr oder weniger gebräuchlichen Terminologie ein sehr wichtiger Faktor ist (vgl. auch Soergel 1999:1120), ohne die sie eine ihrer zentralen Funktionen – die Integration divergierender Terminologien – nicht erfüllen könnten. Ein zweiter Grund für einen Vergleich mit Thesauri ist, dass Thesauri durch ihre Orientierung an Prinzipien der Facettenklassifikation, einer häufig polyhierarchischen Darstellung und v.a. der Darstellung von Beziehungen zwischen Konzepten

¹⁶ Wie bereits erwähnt gibt es auch nicht hierarchische Ontologien, die hier aber nicht berücksichtigt werden.

mehr den Prinzipien und Darstellungszielen von Ontologien entsprechen als Klassifikationen, die nur allgemeine paradigmatische semantische Beziehungen zwischen einzelnen Klassen zulassen (vgl. Reimer:181; Manecke:144).

4.1 Thesauri – Aufbau und Nutzung

Im Folgenden sollen kurz Aufbau und Nutzung von Thesauri skizziert werden, um auf diese Weise Gemeinsamkeiten und Unterschiede von Thesauri und Ontologien beleuchten zu können.

4.1.1 Strukturelle Ebene

Ein Thesaurus wird in der DIN 1463 folgendermaßen definiert:

„Ein Thesaurus im Bereich der Information und Dokumentation ist eine geordnete Zusammenstellung von Begriffen und ihren (vorwiegend natürlichsprachigen) Bezeichnungen, die in einem Dokumentationsgebiet zum Indexieren, Speichern und Wiederauffinden dient.

Er ist durch folgende Merkmale gekennzeichnet:

a) Begriffe und Bezeichnungen werden eindeutig aufeinander bezogen („terminologische Kontrolle“), indem

- Synonyme möglichst vollständig erfasst werden,*
- Homonyme und Polyseme besonders gekennzeichnet werden,*
- für jeden Begriff eine Bezeichnung (Vorzugsbenennung, Begriffsnummer oder Notation) festgelegt wird, die den Begriff eindeutig vertritt,*

b) Beziehungen zwischen den Begriffen (repräsentiert durch ihre Bezeichnungen) werden dargestellt.“

Burkart fügt dem hinzu, dass Thesauri dabei präskriptiv festlegen, welche Begriffe – bezogen auf den jeweiligen Geltungsbereich, und das heißt hier die Indexierung in einem Dokumentationsbereich und nicht die Repräsentation sämtlichen dort vorhandenen Wissens – und welche Bezeichnungen dafür jeweils zugelassen sind (vgl. Burkart:160). Inwieweit diese Einschränkung gerade im Kontext relevanter werdender Volltextrecherche aufrecht erhalten werden kann, bleibt an dieser Stelle dahingestellt. Es scheint jedoch wahrscheinlich, dass diese enge Bindung von Begriffen und Terminologie in Thesauri an die jeweilige Indexierungspraxis zunehmend aufgelöst wird: einerseits zu Gunsten einer Orientierung an der von den (End-)Nutzern tatsächlich gebrauchten Terminologie und andererseits zu Gunsten eines größeren Allgemeinheitsgrades durch Einbindung in größere Zusammenhänge und Systeme (z.B. zur Textanalyse, in größere semantische Netzwerke) (vgl. Milstead). In der von Aitchison u.a. aus der ISO-Norm 2788 übernommenen Thesaurusdefinition wird außerdem betont, dass es sich bei den in Thesauri beschriebenen semantischen Relationen um A-priori-Beziehungen handelt (Aitchison u.a.:1) – d.h., dass sie sich auf die begriffliche Ordnung der Welt in dem jeweiligen Wissensgebiet beziehen und nicht auf die im jeweils beschriebenen Dokument hergestellten Beziehungen.¹⁷

Thesauri verzeichnen also Begriffe (Konzepte) und Bezeichnungen für sie und setzen sie zueinander in Beziehung.

¹⁷ Unterschiede zwischen den einzelnen nationalen und internationalen Normen sollen hier aber nicht weiter berücksichtigt werden.

Tragend für Thesauri sind dabei hierarchische Relationen, v.a. generische IS-A-Relationen, die Vererbungsprinzipien zulassen (vgl. Aitchison u.a.:56), und die Teil-Ganzes-Beziehung, die z.T. ebenfalls als hierarchiebildend angesehen werden kann. Die Teil-Ganzes-Beziehung wird jedoch nicht immer eigens ausgewiesen (vgl. Aitchison u.a.:58f; Burkart:174f). Dieser Relationstyp wird in Thesauri meistens mit „broader term“ (BT) und „narrower term“ (NT) angezeigt. Hierarchische Relationen in Thesauri sind dabei oft polyhierarchisch angelegt (vgl. Aitchison u.a.:60) und lassen damit auch relativ flexible Suchmöglichkeiten zu (vgl. Manecke:143). Individualkonzepte, die mit Eigennamen bezeichnet werden, können mit Hilfe der „Instanz“-Relation (nicht zu verwechseln mit der Instance-Of-Relation in objektorientierten Computersystemen) in einen Thesaurus aufgenommen werden (vgl. Aitchison u.a.:59).

Neben der hierarchischen gibt es in Thesauri die Assoziationsrelation, die auf die eine oder andere Weise zusammengehörige, aber nicht in einem hierarchischen Verhältnis stehende Begriffe kombiniert. Angezeigt wird sie mit „related term“ (RT). Unter diesem Relationstyp werden sehr verschiedene Arten von Beziehungen zusammengefasst, die Vergabe dieser Relation sollte daher überlegt und kontrolliert, ggf. in Abhängigkeit von der Zwecksetzung des Thesaurus erfolgen (vgl. Aitchison u.a.:59ff; Burkart:174). Auch wenn diese Arten von Assoziationsbeziehungen aus Gründen einer kontrollierten Vergabe unterschieden werden, werden sie nicht im Thesaurus ausgewiesen. Allerdings erfordern ein automatisierter Einsatz von Thesauri zunehmend eine Differenzierung dieser Relationen und v.a. ihre konsistente Zuweisung (vgl. Aitchison u.a.:66).

Hier ist noch zu erwähnen, dass je nach Verwendungszweck des Thesaurus der Umfang, in dem assoziative Beziehungen ausgewiesen werden, stark variieren kann. In dem „klassischen“ Thesaurus geht es nicht darum, alle denkbaren Beziehungen, sondern nur die für den jeweiligen geplanten Anwendungsbereich sinnvollen Beziehungen auszuweisen (vgl. Burkart:174). Mit zunehmender Benutzung von Thesauri auch für Volltextrecherchen und von Endnutzern ist es jedoch nötig, einen möglichst großen Teil des spezialisierten Vokabulars, das von Autoren in ihren Texten oder Thesaurusnutzern in Anfragen benutzt werden könnte, zu reflektieren (vgl. Milstead).

Der systematische Aufbau von Thesauri beruht auf den Prinzipien der Klassifikation, und zwar zumeist auf denen der Facettenklassifikation, die ein Hilfsmittel bietet, den jeweils darzustellenden Bereich mit seinen konzeptuellen Beziehungen begrifflich zu analysieren (vgl. Aitchison u.a.:68f). Sie werden ergänzend zu der zentralen hierarchischen Relation eingesetzt. Anders als normalerweise in einer Klassifikation, stehen in Thesauri zum Teil hierarchische Teilketten unverbunden nebeneinander und bilden keine große zusammenhängende Baumstruktur. Dadurch wird die hierarchische Relation deutlich abgeschwächt. Fundamentale Facetten können als generelle Kategorien für die Indexierungsterme dienen, deren „upper level“ sowohl in ANSI-Standards als auch in der ISO-Norm 2788 erfasst ist. Sie werden ebenfalls als Konzeptkategorien im semantischen Netzwerk des UMLS verwendet (vgl. Aitchison u.a.:71).

Facetten werden auch als Rollenindikatoren gebraucht (vgl. Aitchison u.a.:71, 89; Fugmann:104ff). Rollenindikatoren werden – ebenso wie so genannte Links – beim Indexieren wie auch beim Recherchieren eingesetzt, um einzelne Dokumente genauer zu beschreiben, d.h. darzulegen, in welchem Zusammenhang und welchem Sinn ein Term im vorliegenden und zu indexierenden Dokument gebraucht wird (z.B. Tomate als Pflanze oder als End-Produkt, vgl. Aitchison u.a.:71, 89). Ziel ist es, beim Suchvorgang die Treffsicherheit zu erhöhen. Auf diese Weise können einige Mehr-

deutigkeiten von Bezeichnungen aufgelöst werden, Beziehungen zwischen Begriffen können jedoch nicht beschrieben werden (vgl. Reimer:181).

Als weiteres Hilfsmittel, die Precision im Suchergebnis durch eine bessere Dokumentbeschreibung zu erhöhen, kann als spezieller Kode die Art und Weise der Behandlung eines Themas im vorliegenden Dokument angegeben werden (z.B. „theoretisch“, „experimentell“ o.Ä.). Diese Angaben beziehen sich dabei auf das Dokument als Ganzes, nicht auf die Verwendung einzelner Terme darin. Ebenfalls können Schlagwörter gewichtet vergeben werden (vgl. Aitchison u.a.:92ff).¹⁸

4.1.2 Terminologische Ebene

Thesauri sind v.a. kontrollierte Vokabulare. Sie verzeichnen im Allgemeinen Nomina, z.T. auch Verben; Adjektive und Adverbien sind nur in zusammengesetzten Ausdrücken erwünscht, Artikel nur in Eigennamen (vgl. Aitchison u.a.:19ff, DIN 1463). Mögliche Wortkombination, Zusammensetzungen und Schreibweisen sind im Einzelnen festgelegt (vgl. Aitchison u.a.:38ff; Burkart:167). Es erfolgt eine Synonym- und Polysemkontrolle: Da es sich bei einem Thesaurus primär um eine geordnete Zusammenstellung von Begriffen, nicht von Bezeichnungen handelt, werden entsprechend begriffliche Äquivalenzklassen gebildet, in denen Synonyme zusammengefasst sind. Polyseme werden in verschiedene Klassen aufgespalten, häufig durch den Zusatz von Qualifikatoren (vgl. Burkart:165ff). In der Regel wird aus den Elementen einer Äquivalenzklasse eine Vorzugsbenennung gewählt, die als Deskriptor beim Indexieren (und auch bei der Recherche) benutzt werden muss. Die Wahl des Deskriptors richtet sich nach dem im Fachgebiet üblichen Sprachgebrauch, zudem sollte er ein guter Repräsentant des Begriffes sein und möglichst einfach (vgl. Burkart:169). Es gibt jedoch auch Thesauri ohne Vorzugsbenennung, bei denen alle Elemente einer Äquivalenzklasse gleichermaßen verwendet werden können (vgl. Burkart:168).

Definitionen von Begriffen werden in Thesauri im Allgemeinen nicht geliefert, die Begriffe sollten v.a. aus der Position der Bezeichnungen in der Thesaurusstruktur deutlich werden. Definitionen von Deskriptoren, die aber nicht erschöpfend sein müssen, werden zum Teil als Interpretationshilfe angegeben. In Scope Notes werden Einschränkungen für den Anwendungsbereich von Deskriptoren in der Indexierung, Hinweise auf besondere Gebrauchsregularitäten usw. erwähnt (vgl. Aitchison u.a.:34f; DIN 1463).

4.1.3 Nutzung

Thesauri werden in ihrer klassischen Verwendung bei der Indexierung von Dokumenten wie auch bei der Recherche gleichermaßen benutzt, um Dokumenten Schlagwörter zuzuweisen, anhand derer sie wieder aufgefunden werden können. Beides erfolgt in der Regel von geschultem Fachpersonal. Es gibt aber auch reine Indexierungs- wie auch reine Suchthesauri, die sich in ihrer Ausgestaltung – in Abhängigkeit von verschiedenen Einsatzbereichen – v.a. bezüglich der verzeichneten Deskriptoren wie auch des Umfanges der in einer Äquivalenzklasse aufgenommenen Synonyme unterscheiden.

¹⁸ Weitere Hilfsmittel, die Precision zu erhöhen, sind ausschließlich bei der Formulierung der Suchfrage anzuwenden (z.B. Boole'sche Operatoren, Distanzoperatoren, Stemming und Trunkierung), ich fasse sie daher hier nicht als strukturelle Möglichkeiten eines Thesaurus auf, bestimmte Beziehungen zwischen Begriffen darzustellen.

Indexierungsthesauri werden v.a. für Datenbanken benutzt, deren meist nicht professionelle Endnutzer Thesauri nicht nutzen, der Thesaurus wird im Hintergrund eingesetzt. Suchthesauri werden zumeist von ebenfalls nicht professionellen Endnutzern als Hilfsmittel für Volltextdatenbanken genutzt (vgl. Aitchison u.a.:1ff). Prinzipiell sind Thesauri bereichs- und adressatenspezifisch; davon hängen das ausgewählte Vokabular, die Spezifität des Thesaurus und auch die aufgeführten Relationen ab.

Dadurch, dass Thesauri im Allgemeinen auch elektronisch zur Verfügung stehen, ergeben sich zahlreiche, zum Teil neue Nutzungsmöglichkeiten: von der (automatischen) Anfrageerweiterung über hierarchisch untergeordnete Begriffe über Möglichkeiten wie Browsing (vgl. Aitchison u.a.:3) zur Bereitstellung weiterer Suchbegriffe, einfacheres Wechseln zwischen verschiedenen strukturierten Ansichten, Einsatz in (halb-)automatischer Indexierung, Unterstützung im Volltextretrieval durch Gewichtung von Suchwörtern (vgl. Aitchison u.a.:82; Köpcke 2002:325; Milstead). Thesauri werden zunehmend auch als Hilfsmittel für die Internetsuche eingesetzt, sie sind häufig nicht oder nur auf Anforderung für den Nutzer unmittelbar zugänglich. Oft arbeiten sie ausschließlich im Hintergrund und stellen dort zusätzliche Suchbegriffe, v.a. Synonyme und Hyponyme, zur Verfügung (vgl. Aitchison u.a.:2f; Milstead).

Für die zukünftige Entwicklung von Thesauri, mit der sie sich an neue Anwendungsfelder und -bedingungen in zunehmend elektronischen Umgebungen anpassen sollen, wird ihre Integration in größere (semantische) Netzwerke gesehen (vgl. Milstead; Soergel 1996), in denen Thesauri die „semantische Komponente“ darstellen sollen. Es gibt zunehmend Vorschläge, mehr und feiner differenzierte Relationen zwischen Thesauri zuzulassen und diese ihrerseits einer Kategorisierung und terminologischen Kontrolle zu unterwerfen (vgl. Aitchison u.a.:62f). In diesem Zusammenhang ist auch das UMLS zu sehen, das bei einem Vergleich von Thesauri und Ontologien eine interessante Zwitterstellung einnimmt. Auch die Rolle von Definitionen, für Ontologien ein zentrales Element, ändert sich: Definitionen wandeln sich von ursprünglich zusätzlichen Informationen zum besseren Gebrauch zu integralen Bestandteilen von Thesauri, es werden mehr Definitionen aufgenommen und auch eine formale logische Form wird diskutiert (vgl. Aitchison u.a.:34).

4.2 Der Nutzen von Ontologien

Thesauri und Ontologien sind vergleichbare, strukturell sehr ähnliche Instrumente, die zu inhaltlicher Erschließung eingesetzt werden können. Beide verzeichnen Konzepte menschlichen, v.a. wissenschaftlichen Denkens, die sie (meistens) in einer Abstraktionshierarchie organisieren, darüber hinaus werden die Konzepte in weitere, mehr oder weniger differenzierte Beziehungen zueinander gesetzt. Beide bemühen sich um einen terminologischen bzw. sprachlichen Zugang zu diesen Konzepten. Ziel ist, menschliches Wissen zu repräsentieren. Thesauri sind dabei explizit an der Wissenschaft in den einzelnen Bereichen orientiert, Ontologien sind dies in vielen Fällen auch. Allerdings gibt es auch Ontologien wie Cyc, die v.a. an Alltagswissen orientiert sind, oder Ontologien für ganz andere Einsatzbereiche wie Prozessmodellierung oder Arbeitsabläufe in Organisationen, die eine andere Gegenstandsmodellierung erfordern. Thesauri und Ontologien bemühen sich, sowohl die Konzepte als auch die Terminologie, die in dem entsprechenden Bereich üblich sind, zu reflektieren.

Wesentliche Unterschiede werden zum einen im Formalisierungsgrad gesehen, der sich v.a. auf die Definition der Konzepte und Relationstypen zwischen ihnen bezieht und der bei Ontologien meistens höher ist als bei Thesauri, zum anderen in der

Reichhaltigkeit und Differenziertheit der Relationen zwischen den Konzepten, mit der die Darstellung auch von dokumentspezifischen Beziehungen ermöglicht werden soll (vgl. Köpcke 2002:326). Dazu kommt noch der entscheidende Unterschied, Schluss- und Inferenzregeln anwenden zu können, was allerdings auch je nach Ontologie sehr verschieden ausgeprägt ist.

Mit Formalisierung und Beziehungsreichtum versuchen Ontologien einzulösen, was in der Literatur auch als zeitgemäße Ausstattung von Thesauri angesehen wird, die einen Einsatz in digitalen Umgebungen, bei der Arbeit mit Volltexten etc. ermöglichen soll – in der Regel allerdings, ohne dass die Ersteller von Ontologien davon wüssten oder dies zumindest zur Kenntnis nähmen (vgl. Soergel 1999:1120; Vickery:284).

4.2.1 Strukturelle Ebene

Betrachtet man die Konzepte und Relationen von Thesauri und Ontologien, so stellt man erstaunlich viele Parallelen fest: Allerdings ist ein Vergleich in diesem Allgemeingrad ziemlich schwierig, da es gerade bei Ontologien große Unterschiede gibt – nicht zuletzt, weil in dieser neueren Entwicklung Standardisierungsprozesse noch im Gange sind. Dies ist nicht nur im Hinblick auf die genaue Ausgestaltung und Formalisierung der Konzepte und Relationstypen der Fall, sondern auch hinsichtlich ihrer Anwendungsbereiche, Zielsetzungen und dem Anspruch an Umfang und eigener Erklärungskraft.

Die meisten Ontologien beruhen zentral auf der hierarchischen IS-A-Beziehung, die der hierarchischen Abstraktionsrelation entspricht, die in Thesauri die zentrale klassifikatorische Hierarchie bildet. Allerdings wird dieser Relationstyp in Thesauri oft mit der hierarchischen partitiven Relation vermischt benutzt, kann aber auch getrennt ausgewiesen werden. Andere Relationen werden in Thesauri als „assoziiert“ zusammengefasst – auch wenn der Einsatz v.a. beim Indexieren kontrolliert und reflektiert erfolgen soll (vgl. Aitchison u.a.:60ff). Hier versprechen Ontologien mit der einfachen Erweiterbarkeit durch weitere Relationstypen tatsächlich eine Bereicherung, insbesondere bei bereichsspezifischen Anwendungen wie z.B. in der Medizin, da in diesen Ontologien „maßgeschneiderte“ Relationstypen gestaltet werden können.

Ontologien organisieren ihre Hierarchien z.T. entlang einiger grundlegender Kategorien oder auch Dichotomien (z.B. konkret vs. abstrakt), denen die jeweils untergeordneten Konzepte hierarchisch zugeordnet sind. Diese grundlegenden Kategorien sind zentral für die Kompatibilität und modulare Nutzbarkeit verschiedener Ontologien, daher sind sie auch Gegenstand von Standardisierungsbestrebungen. Es gibt einige allgemeine Kategorien, die in vielen Ontologien unabhängig von dem jeweiligen Gegenstandsbereich auftauchen, darunter z.B. „things“, „processes“, „events“ u.a. Auch „time“ spielt eine zentrale Rolle (vgl. dazu Fridman Noy/Hafner:68f).

Dies erinnert stark an die in Thesauri benutzten generellen Kategorien, die der ISO-Standard in „concrete entities“, „abstract concepts“ und „individual entities“ einteilt, sowie an manche fundamentale Facetten (z.B. „things“, „actions“, „time“, vgl. Aitchison u.a.:17f,70f).

In Bezug auf solche generellen, grundlegenden Kategorien wäre genauer zu untersuchen, inwieweit Ontologien und Thesauri hier tatsächlich strukturelle Ähnlichkeiten und Parallelen in der Verwendung aufweisen. Für Thesauri existieren für diese Ebenen bereits Standards, es wäre nahe liegend, dass Ontologien und Thesauri hier zumindest von einander profitieren könnten.

Auch die jeweiligen Konstruktionsprinzipien¹⁹ sind durchaus vergleichbar, bei Verfahren der Ontologieerstellung wird auf u.a. solche der Thesauruserstellung zurückgegriffen, wie z.B. Expertenwissen, Literaturanalyse und der Rückgriff auf einschlägige Wörterverzeichnisse (vgl. Köpcke 2001:24).

Die Formalisierung und damit die maschinelle Verarbeitbarkeit von Ontologien ist ebenfalls als Vorteil anzusehen. In Ontologien sind Definitionen von Konzepten und auch von Relationstypen eigentlich erforderlich, in der konkreten Umsetzung finden sich aber deutliche Unterschiede. WordNet z.B., das als wenig formale Ontologie gilt, führt seine Definitionen ebenfalls nur als Veranschaulichung für Nutzer mit, wie dies auch Thesauri tun, Cyc dagegen legt als sehr formale Ontologie großen Wert auf hochgradig formale Definitionen.

Ein wesentlicher Unterschied zwischen Ontologie und Thesaurus ist sicherlich in der Existenz von Inferenzregeln und Schlussmechanismen zu sehen, über die Ontologien verfügen und die es ihnen ermöglichen, über Informationen Aussagen zu machen. Dies kann prinzipiell genutzt werden z.B. zum automatischen Indexieren, zur besseren Volltextsuche, aber auch zur Konsistenzprüfung von Systemen. Inwieweit dies jedoch in allen Ontologien vorhanden ist und wie ausgefeilt diese Regeln jeweils sind, bleibt dahingestellt. Das UMLS verfügt z.B., soweit ich sehe, über keine nennenswerten Schlussmechanismen. Hier sei jedoch noch einmal auf die starke Heterogenität verwiesen, die Aussagen auf diesem Allgemeingrad oft schwierig macht.

Ein kurzer Blick unter dem Aspekt der Leistungsfähigkeit von Ontologien im Vergleich zu Thesauri soll hier noch auf einige Anwendungen von Ontologien in Projekten digitaler Bibliotheken geworfen werden, die ich in Kapitel 3.2 bereits erwähnt habe.

Alle als Beispiel angeführten Ontologien sind hierarchisch organisiert, wobei die Hierarchie eine unterschiedliche Rolle spielt. In der UMDL (Weinstein/Birmingham) z.B. ist sie zentral, das „Rückgrat“ der Ontologie, entsprechend relevant ist die exakte Ausgestaltung der hierarchischen Relationen (z.B. der dort angesiedelten Vererbungsmechanismen) und ihre Rolle für die Erfüllung der Aufgabe der Ontologie. Das ScholOnto-Projekt (Buckingham Shum u.a.) dagegen konzentriert sich mehr auf die Relationen zwischen Werken, Autoren etc., die für die Darstellung der argumentativen Beziehungen relevanter sind.

Die hier angeführten Ontologien machen sich die Möglichkeit, mehr und reichhaltigere semantische Relationen zu benutzen, durchaus zu Nutze, auf ihnen basieren in der Regel auch die verwendeten Schlussmechanismen, die sich v.a. auf die Beantwortung von Suchfragen beziehen. Die Antworten werden aus Konzepten und ihren Beziehungen heraus generiert, über Inferenzen werden Anfragen erweitert. In ScholOnto sollen Schlussmechanismen benutzt werden, um gedankliche „Schulen“ aus den Autorenbeiträgen zu generieren; die UMDL setzt sie bei der Konversion aus MARC-Daten ein. Auch die in der Systemarchitektur „versteckten“ Anwendungen nutzen die Schlussmechanismen z.B. zur Konsistenzprüfung (Stuckenschmidt/van Harmelen) oder für die Einbettung in Agentensysteme (Endres-Niggemeyer).

Es fällt auf, dass sich die meisten Ontologien gerade im Zusammenhang mit der Verbesserung von Recherchemöglichkeiten nicht auf die Sacherschließung im eigentlichen Sinne beziehen, dem ureigenen Gebiet der Thesaurusanwendung. In den Beispielen geht es um bibliografische Relationen zwischen verschiedenen Veröffent-

¹⁹ Auf die jeweiligen Konstruktionsprinzipien gehe ich hier allerdings nicht detaillierter ein (vgl. aber Köpcke 2001:13ff, 24ff).

lichungen (Weinstein/ Birmingham) bzw. um argumentative Zusammenhänge zwischen Veröffentlichungen (Buckingham Shum u.a.). Auch Welty/Jenkins, die sich v.a. mit thematischer Klassifikation befassen, also mit Sacherschließung im besten Sinne, konzentrieren sich auf die Problematik, wie verschiedenartige Gegenstände einheitlich in einer Klassifikation klassifiziert werden können und wie die Klassen und Relationstypen dafür im Detail gestaltet sein müssen. Es geht hier nicht um die sachlichen Beziehungen im thematischen Gebiet. Diese Aspekte, die in den Ontologie-Projekten aufgegriffen wurden, werden im Allgemeinen von Thesauri nur sehr am Rande berücksichtigt. So gibt es z.B. zwar Möglichkeiten, die Darstellungsweise von Dokumenten (z.B. theoretisch, experimentell etc., vgl. Kapitel 4.1.1) in Thesauri zu verzeichnen, dies aber nur beiläufig und auf einer sehr allgemeinen Ebene. Die gegenseitigen Bezüge könnten hier nicht verdeutlicht werden.

Ontologien als Mittel zur Strukturierung (des Wissens) eines Sachgebietes werden zum einen im Zusammenhang größerer Softwaresysteme eingesetzt (ein Beispiel ist das automatische Zusammenfassen im Kontext mit Verfahren des Natural Language Understanding, vgl. Endres-Niggemeyer), andererseits zur Integration verschiedenartiger Metadaten und Thesauri (z.B. Amann u.a.; UMLS). Bei dem Einsatz in einem Mediationssystem (Amann u.a.) allerdings wird die sachliche Strukturierung von einem Thesaurus geliefert; das UMLS wird vielfach selbst als Thesaurus gesehen und bezieht auch einen Großteil seiner Datenstruktur aus den integrierten Thesauri. Die Affinität zum Thesaurus scheint hier besonders groß zu sein, es finden sich fließende Übergänge.

4.2.2 Terminologische Ebene

Interessant ist auch die Frage nach dem Umgang mit Terminologie und Vokabularkontrolle bei Ontologien. Einerseits wird bei Ontologien großer Wert darauf gelegt, dass die in ihnen repräsentierten Einheiten Konzepte und nicht Wörter sind, andererseits werden Ontologien durchaus auch als „Wörterbücher“ bezeichnet. Eines der zentralen Ziele beim Einsatz von Ontologien ist die Integration von Terminologie, was sich – natürlich – auf die Integration verschiedener Bezeichnungskonventionen bezieht, die auf ein gemeinsames Konzept zurückgeführt werden sollen. Dies tun auch Thesauri im Rahmen der terminologischen Kontrolle: Synonyme werden auf ein gemeinsames Konzept zusammengeführt, Polyseme werden nach den verschiedenen Konzepten differenziert. Eine Integration von verschiedenen Terminologien (oder auch Thesauri) muss hier auch an der Bezeichnungsebene ansetzen, auch wenn ebenfalls eine Integration der Konzeptstrukturen erforderlich ist.

Thesauri gelten als präskriptiv, da sie vorschreiben, welche Begriffe für die Beschreibung des Gegenstandsbereiches zur Verfügung stehen und welche Bezeichnung für einen Begriff benutzt werden muss. Diese Präskriptivität bezieht sich auf der terminologischen Ebene jedoch ausschließlich auf den Zweck des Thesaurus: das Indexieren. Nur hierfür gilt, dass eine bestimmte Bezeichnung verwendet werden muss, um einen bestimmten Begriff auszudrücken.²⁰ In der Auswahl der Deskriptoren wie auch der relevanten Begriffe orientieren sich Thesauri wie auch Ontologien am allgemeinen bzw. im Fachgebiet üblichen Sprachgebrauch und berücksichtigen auch die von Nutzern in Anfragen wahrscheinlich verwendeten Formulierungen. Ein Thesaurus kann in dieser Hinsicht als „common agreement“ betrachtet werden (vgl. Amann u.a.:225). In dieser Hinsicht sind sie also eher um Deskriptivität bemüht.

²⁰ Dies gilt so natürlich nur für Thesauri, die mit Vorzugsbenennungen arbeiten.

Vor einer Präskriptivität auf der begrifflichen Ebene sind m.E. auch Ontologien nicht gefeit. Obwohl der Anspruch oft lautet, das Wissen in einem Bereich zu repräsentieren, indem auch verschiedene Perspektiven integriert, d.h. vermutlich auch berücksichtigt werden sollen, birgt meiner Ansicht nach gerade die Rückführung auf gemeinsame Konzepte und Relationstypen wie auch deren strenge, formale Definition ein großes Potenzial an präskriptiver Wirkung. Aus der hier vorliegenden Literatur wird nicht ersichtlich, wie z.B. mit neuen wissenschaftlichen Ansätzen, mit Nuancen in Konzeptualisierungen, die vom „Commonsense“ der wissenschaftlichen Gemeinschaft abweichen, mit Lehrmeinungen und ihrer Durchsetzung gerade in verbindlichen Definitionen umgegangen wird. Es ist zu prüfen, ob – zumindest mit Fachontologien – nicht ein Instrument geschaffen wird, mit dem ein Kanon anerkannter wissenschaftlicher Konzepte etabliert wird, der davon abweichende nicht erfassen und auch nicht beschreiben kann. Dies würde allerdings den Zielsetzungen von Ontologien – Repräsentation von Wissen, Integration verschiedener Perspektiven und Bereitstellen von elektronisch vorliegenden Informationen verschiedenster Art – widersprechen. Selbstverständlich müssen auch Ontologien laufend gepflegt werden, um nicht an Aktualität zu verlieren. Wie Buckingham Shum u.a. bemerken, kann das, wie auch die Festschreibung bestimmter Perspektiven, ein Problem werden (vgl. Buckingham Shum u.a.:240).

In jedem Fall legen Thesauri einen deutlich größeren Schwerpunkt auf die Sammlung und Kontrolle von Wortgut, auch wenn dieses in vielen Ontologien eine wichtige Rolle spielt.²¹ Um Verzeichnisse von Konzepten handelt es sich in beiden Fällen. Es sei an dieser Stelle nur noch darauf hingewiesen, dass Ontologien im Allgemeinen offensichtlich von einer prinzipiellen Sprachunabhängigkeit von Konzepten ausgehen, was ihnen die Annahme einheitlicher Konzepte auch über verschiedene Terminologien, sogar Sprachen hinweg ermöglicht. Im Gegensatz dazu erfolgt die Rückführung von Synonymen auf gemeinsame Konzepte im Thesaurus ausgesprochen zweckgebunden für das Information Retrieval, eine tatsächliche – ontologische – Identität wird hier nicht behauptet.

4.2.3 Nutzung

Das klassische Anwendungsgebiet von Thesauri, die Bereitstellung von standardisiertem und kontrolliertem Vokabular zur Indexierung von Dokumenten und ihrer Recherche in Datenbanken mit Hilfe Boole'scher Suchverfahren (vgl. Milstead) ist sicherlich kein geeignetes Anwendungsgebiet für Ontologien. Die Nutzung in neuen Informationssystemen wie z.B. digitalen Bibliotheken stellt neue Anforderungen an Thesauri, denen sie sich mit Weiterentwicklung auf verschiedenen Gebieten zu stellen suchen. Dadurch, dass Thesauri meistens auch in elektronischer Form vorliegen, haben sich neue Nutzungsbereiche eröffnet. Die Anwendung erfolgt zunehmend „versteckt“ in Form von Anfrageerweiterung über Synonymen-Cluster, die für einen der im Cluster enthaltenen Terme stehen (Äquivalenzklasse) (vgl. Aitchison u.a.:83), sie unterstützen (halb-) automatisch Such- und Indexierungsprozesse (vgl. Aitchison u.a.:2f) wie auch Volltextsuche (vgl. Milstead), Suche durch End-Nutzer in Inter- und Intranets (vgl. Aitchison u.a.:2; Dextre-Clarke; Milstead). Es wird hoher Bedarf ange-

²¹ Hier ist es aber sicherlich notwendig, zwischen Ontologien mit verschiedenen Verwendungszwecken zu unterscheiden: Eine große, allgemeine Ontologie wie Cyc, die v.a. menschliche Schlussverfahren imitieren möchte, legt sicherlich weniger Wert auf Terminologiearbeit als eine fachbezogene Ontologie, die v.a. Fachwissen (und -termini) zu einander in Beziehung setzen möchte. Eine Ontologie zur Abbildung von Abläufen in Unternehmen ist hier noch einmal ganz anders einzuordnen.

meldet, die Vereinbarkeit und Integration von Thesauri voranzutreiben, mit Zielen wie Suchmöglichkeiten in verschiedenen Sprachen, über sehr heterogene Datenquellen hinweg, Interkulturalität etc. Als Methoden der Integration und des Vereinbar-Machens werden Mapping, Switching, Zwischensprachen und Makrothesauri genannt, die automatisch und im Hintergrund operieren können (vgl. Aitchison u.a.:174ff).

Diese für die Thesaurusentwicklung gesteckten Ziele – Integration verschiedener Datenbanken und Terminologien über Sprachen und Kulturen hinweg, versteckte Nutzung im Hintergrund, semantische Reichhaltigkeit und formale Definition zur besseren Dokumentbeschreibung und zur Verbesserung der Recherchemöglichkeiten – reklamieren Ontologien ebenfalls für sich.

4.3 Fazit

Thesauri und Ontologien sind strukturell sehr ähnliche Instrumente. Die hier betrachteten Thesauri sind explizit zur dokumentarischen Inhaltserschließung und Recherche entworfen. Ontologien haben einen größeren Anwendungsbereich, zentral ist aber auch hier die Erschließung und Bereitstellung von Informationen. Beide haben die Repräsentation von Wissen in Form von Konzepten und Beziehungen zwischen ihnen zum Ziel. Hinsichtlich der Formalisierung, die eine maschinelle Verarbeitung ermöglicht, und der Reichhaltigkeit von darstellbaren semantischen Relationen sind Ontologien Thesauri überlegen, auch kommen in Ontologien automatische Schlussverfahren zum Einsatz.

Diese Formalisierung und bessere Differenzierung von semantischen Beziehungen wie auch Schlussmechanismen werden jedoch auch für eine Anpassung von Thesauri an die Gegebenheiten in neueren Informationssystemen, wie z.B. digitale Bibliotheken, zunehmend gefordert, die (angestrebten) Einsatzbereiche von Thesauri sind mit denen von Ontologien vergleichbar: Integration von heterogenen Terminologien, Einsatz in semantischen Rechercheinstrumenten, bessere Aufbereitung und Bereitstellung von digitalen Volltexten.

In dieser Hinsicht können Ontologien als Weiterentwicklung von Thesauri betrachtet werden, ohne dass jedoch in der Ontologienentwicklung auf informations- oder bibliothekswissenschaftliche Grundlagen Bezug genommen würde.

Auch die konkreten Anwendungen von Ontologien in Projekten (digitaler Bibliotheken) lassen sich als Weiterentwicklung von Thesauri unter Nutzung der neuen technischen Möglichkeiten begreifen. Ontologien werden in größeren semantischen Netzwerken eingesetzt, liefern hier z.T. die semantische Komponente (Stuckenschmidt/van Harmelen; Endres-Niggemeyer) und arbeiten versteckt in Mediations- und Agentensystemen. Zur Verbesserung der Recherche an der Oberfläche werden in den hier angeführten Beispielen allerdings v.a. solche Ontologien angewendet, die andere Zugänge eröffnen sollen als herkömmliche Sachrecherche, oder die Zusammenhänge möglich machen sollen, die in Thesauri nur mangelhaft repräsentiert sind.

Der „Mehrwert“ von Ontologien gegenüber Thesauri liegt auch hier in ihrer Formalisierung und damit einhergehenden Einsetzbarkeit in komplexeren Softwaresystemen, verbunden mit dem Einsatz von Schlussmechanismen. Ein weiterer Vorteil liegt in der reichen Gestaltbarkeit von Beziehungen zwischen Konzepten – die sich in den konkreten Anwendungen allerdings häufig andere Bereiche beziehen als Thesauri. Inwieweit dies den praktischen Bedürfnissen und Nutzungsweisen in digitalen Bibliotheken entspricht, bleibt hier dahingestellt.

Exklusiv für Ontologien sind die in den Beispielen dargestellten Herangehensweisen allerdings nicht, auch diese werden allgemein für moderne Zugänglichkeit und Recherchemöglichkeiten von Informationen gefordert (vgl. Aitchison u.a.:174; Soergel 1996).

5 Fazit

Ontologien stellen formale Repräsentation von Wissen über einen Ausschnitt der Welt dar. Dieses Wissen wird in einem Begriffssystem dargestellt, dessen Begriffe wie auch Beziehungen zwischen Begriffen definiert sind und das in der Lage ist, Inferenzen zu generieren. Es sollen die verschiedenen Perspektiven auf diesen Weltausschnitt wie auch die verschiedenen terminologischen Zugänge zu diesem Weltausschnitt reflektiert und integriert werden.

Ein solches Begriffssystem kann in sehr verschiedenen Bereichen eingesetzt werden, von der Objektmodellierung bis zur Verbesserung innerbetrieblicher Kommunikation. Ein zentraler Aspekt in der Anwendung von Ontologien ist der Austausch von Daten in modernen Kommunikationsnetzen wie z.B. dem WWW. Da Ontologien über „Semantik“ verfügen, können sie eingesetzt werden, um auch eine inhaltliche, intelligente Bearbeitung und Integration der ausgetauschten Daten vorzunehmen.

Dies macht sie interessant für digitale Bibliothek, die angesichts der steigenden Informationsflut zum einen große Mengen an Volltexten zu speichern und zugänglich zu machen haben, andererseits ihren Nutzern einen möglichst einheitlichen Zugang zu sehr verschiedenen, heterogenen Datenbeständen ermöglichen sollen, die nicht alle von ihnen selbst verwaltet werden.

Hier können formale Ontologien für digitalen Bibliotheken durchaus von Nutzen sein: Eingebettet in komplexe Systemarchitekturen, integriert in Agenten- und Mediationssysteme können sie helfen, große Mengen digitaler Dokumente zu strukturieren und zugänglich zu machen. Sie sind von Nutzen bei der Integration von verschiedener Terminologie sowohl bei der Suche über verschiedene Datenbanken als auch bei der Recherche in WWW-Seiten, Metadaten und Volltexten.

Sie können genutzt werden, um bei der Informationsrecherche Zusammenhänge in Sachgebieten zu strukturieren und über sie Auskunft zu geben oder Beziehungen zwischen verschiedenen Veröffentlichungen zu verdeutlichen. Im Zusammenhang mit Verfahren des Natural Language Understanding können sie bei Volltextsuche und (halb-)automatischen Erschließungsverfahren eingesetzt werden. Auch in der Analyse von Nutzeranfragen können sie eingesetzt werden.

Gerade in Bezug auf Inhaltserschließung und Recherche drängt sich ein Vergleich mit herkömmlichen Mitteln der Inhaltserschließung wie Thesaurus und Klassifikation auf. Ontologien weisen gerade mit Thesauri viele strukturelle Gemeinsamkeiten auf, sind allerdings der Nutzung in elektronischen Systemen angepasst. Die Unterschiede, die sich zeigen (Beziehungsreichtum, maschinelle Verarbeitbarkeit) wie auch die für digitale Bibliotheken in Frage kommenden Nutzungsweisen (z.B. Suche in heterogenen Datenbeständen, Volltexten und Metadaten) werden auch als Weiterentwicklung und (zukünftige) Einsatzgebiete von Thesauri gesehen. Dass sie hier auf eine reiche Tradition der Begriffsanalyse zurückgreifen können, ist den Erstellern von Ontologien allerdings nur selten bewusst.

Für eine Beurteilung des Nutzens, den ein Einsatz von formalen Ontologien in digitalen Bibliotheken haben kann, ist jedoch eine genaue Beschreibung der jeweiligen bibliothekarischen Anforderungen und ggf. eine Anpassung der Ontologie daran nötig. Hier ist nicht nur zu bedenken, dass es verschiedene Arten von Ontologien mit unterschiedlichen Zielsetzungen gibt, die sie sich nur zum Teil für einen Einsatz in digitalen Bibliotheken anbieten. Auch die Vorteile von Ontologien, Beziehungsreichtum und Formalisierung, müssen auf das konkrete Einsatzgebiet mit seinen Anforderungen bezogen werden. Es kann z.B. differenziert werden zwischen der Nutzung

von Ontologien durch andere Software und der Nutzung durch Menschen. Bei der Nutzung durch Software sind Formalisierung und Beziehungsreichtum sicherlich von Vorteil, auch eine Übersetzung von Anfragen der Nutzer in eine passende Formulierung der Suchfrage kann von Vorteil sein.

Inwieweit diese Vorteile bei dem Gebrauch von Ontologien durch Menschen praktikabel und vorteilhaft sind, ist zu überprüfen. Höhere Exaktheit und Formalität führt nicht unbedingt zu besserem Verständnis durch Menschen, auch die Vielfalt von Relationstypen kann verwirrend sein und zu Problemen einer konsistenten Anwendung führen, v.a. bei einer intellektuellen Erschließung. Ob und unter welchen Bedingungen Ontologien als „Endnutzerontologien“ einsetzbar sind, sollte überprüft werden.

Die von Ontologien gebotene Beziehungsvielfalt kann aber auch von besonderem Wert sein.

Außerdem sollten die von Ontologien eingesetzten Inferenzregeln auf die für das Information Retrieval zentralen Qualitätsmaße Recall und Precision bezogen werden: Gerade z.B. Methoden der Anfrageerweiterung erhöhen zwar den Recall, gehen aber in der Regel auf Kosten der Precision. Hier ist zu überlegen, wie dies bei inferierten, evtl. unsicheren Ergebnissen gestaltet werden soll.

Umgekehrt liegt der Anspruch neuerer Formen der Wissensrepräsentation meisten darin, im Gegensatz zum Thesaurus dokumentenspezifische Relationen darstellen zu können und damit eine genauere Dokumentbeschreibung und präzisere Recherchemöglichkeit zu liefern. Hier ist zu prüfen, ob eine Bereichsontologie z.B. die dafür nötigen Beziehungen und Konzepte tatsächlich bereitstellt, wie z.B. mit neuen, falschen oder vom auch nur vom „Commonsense“ abweichenden Aussagen verfahren wird. Es besteht die Gefahr der Duplizierung des Dokumentes.

Schließlich sei an dieser Stelle noch darauf hingewiesen, dass die Entwicklung und Pflege von Ontologien hohe Kosten mit sich bringt, die natürlich in ein Verhältnis zu dem Nutzen gesetzt werden müssen.

Um Ontologien in digitalen Bibliotheken wirklich gut nutzen zu können, sind hier genaue Anforderungen zu formulieren, an die Ontologien ggf. angepasst werden müssen. Davon konnten hier nur einige angesprochen werden. Bei einer entsprechenden Rückkopplung auch an bisherige dokumentarische Verfahren stellen sie jedoch ein viel versprechendes Werkzeug dar, um mit Anforderungen, die an digitale Bibliotheken gestellt werden, angemessen umzugehen.

6 Abkürzungsverzeichnis

ANSI	American National Standards Institute
DIN	Deutsches Institut für Normung e.V.
DTD	Document Type Definiton
HTML	HyperText Markup Language
IFLA	International Federation of Library Associations and Institutions
IR	Information Retrieval
ISO	International Organization of Standardization
KI	künstliche Intelligenz
MARC	Machine-Readable Cataloguing
MeSH	Medical Subject Headings
NLM	National Library of Medicine (USA)
RDF	Resource Description Framework
SGML	Standard Generalized Markup Language
UMDL	University of Michigan Digital Library
UMLS	Unified Medical Language System
WWW	World Wide Web
XML	eXtensible Markup Language
XSL	eXtensible Stylesheet Language

7 Literatur

- Aitchison, J.; Gilchrist, A.; Bawden, D.: Thesaurus construction and use: a practical manual. London, 2000 (4. Aufl.)
- Amann, B.; Fundulaki, I.; Scholl, M.: Integrating ontologies and thesauri for RDF schema creation and metadata querying. In: International Journal on Digital Libraries 3, 2000, S. 221–236
- Arms, W.Y.: Digital Libraries. Cambridge, Mass. u.a., 2001 (2. Aufl.)
- Buckingham Shum, S.; Motta, E.; Domingue, J.: ScholOnto: an ontology-based digital library server for research documents and discourse. In: International Journal on Digital Libraries 3, 2000, S. 237–248
- Burkart, M.: Thesaurus. In: Buder, M.; Rehfeld, W.; Seeger, T.; Strauch, D. (Hgg.): Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Informationsarbeit. Begründet von Klaus Laisiepen, Ernst Lutterbeck und Karl-Heinrich Meyer-Uhlenried. 4. völlig neu gefasste Ausgabe. München u.a., 1997, S. 160–179
- Chandrasekaran, B.; Josephson, J. B.; Benjamins, V. R.: What are ontologies, and why do we need them? In: IEEE Intelligent Systems, 1999, S. 20–26
- Cyc Applications: <http://www.cyc.com/applications.html> (Zugriff: 18.3.2003)
- Cyc: <http://www.cyc.com/products2.html> (Zugriff: 18.3.2003)
- Dextre Clarke, S.: Thesauri, Topics and Other Structures in Knowledge Management Software. In: Beghtol, C.; Horwarth, L.C.; Williamson, N.J. (Hgg.): Dynamism and Stability in Knowledge Organization. Proceedings of the Sixth International ISKO Conference 10–13 July 2000. Würzburg, 2000, S. 41–47
- DIN 1463, Teil 1: Erstellung und Weiterentwicklung von Thesauri. Einsprachige Thesauri. Berlin, 1987
- Ding, Ying: A review of ontologies with the Semantic Web in view. In: Journal of Information Science 27 (6), 2001, S. 377–384
- Endres, A.; Fellner, D.W.: Digitale Bibliotheken. Informatik-Lösungen für globale Wissensmärkte. Heidelberg, 2000
- Endres-Niggemeyer, B.: Bessere Information durch Zusammenfassen aus dem WWW. 1999. URL: <http://www.ik.fh-hannover.de/ik/person/ben/WEBZusammenfassen.pdf> (Zugriff: 13.4.2003)
- Fellbaum, Chr.: Introduction. In: Dies. (Hg.): WordNet. An electronic lexical database. Cambridge, Mass. u.a., 1998, S. 1–19
- Fensel, D.: Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Berlin u.a., 2000
- Fridman Noy, N.; Hafner, C.D.: The State of the Art in Ontology design. In: AI Magazine 18 (3), 1997, S. 53–74
- Fugmann, R.: Theoretische Grundlagen der Indexierungspraxis. (= Fortschritte in der Wissensorganisation, Bd. 1) Frankfurt a.M., 1992
- Gruber, T.: What is an Ontology? 1993. URL: <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html> (Zugriff: 13.4.2003)

- Köpcke, A.: Ontologien: inhaltliche Erschließung in elektronischen Umgebungen. Diplomarbeit, FH Hannover (FB Bibliothekswesen), 2001
- Köpcke, A.: Ontologien – inhaltliche Erschließung in elektronischen Umgebungen. In: Schmidt, R. (Hg.): Content in Context. Perspektiven der Informationsdienstleistung. Proceedings der 24. Online-Tagung der DGI. Frankfurt a.M., 2002, S. 323–339
- Lenat, D.B.; Guha, R.V.: Building large knowledge-based systems. Representation and Inference in the Cyc-Project. Reading, Mass. u.a., 1990
- Manecke, H.-J.: Klassifikation. In: Buder, M.; Rehfeld, W.; Seeger, T.; Strauch, D. (Hgg.): Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Informationsarbeit. Begründet von Klaus Laisiepen, Ernst Lutterbeck und Karl-Heinrich Meyer-Uhlenried. 4. völlig neu gefasste Ausgabe. München u.a., 1997, S. 141–159
- Miller, G.A. u.a.: WordNet. An On-Line Lexical Database. In: International Journal of Lexicography 3 (4), 1990, S. 235–312
- Milstead, J.: Use of Thesauri in the Full-Text Environment. 1998. URL: <http://www.jelem.com/useof.htm> (Zugriff: 13.4.2003)
- NLM Indexing Initiative: <http://ii.nlm.nih.gov/> (Zugriff: 25.4.2003)
- Opencyc: <http://www.opencyc.org/> (Zugriff: 2.4.03)
- Reimer, U.: Neue Formen der Wissensrepräsentation. In: Buder, M.; Rehfeld, W.; Seeger, T.; Strauch, D. (Hgg.): Grundlagen der praktischen Information und Dokumentation. Ein Handbuch zur Einführung in die fachliche Informationsarbeit. Begründet von Klaus Laisiepen, Ernst Lutterbeck und Karl-Heinrich Meyer-Uhlenried. 4. völlig neu gefasste Ausgabe. München u.a., 1997, S. 180–207
- Rusch-Feja, D.: Digital Libraries – Informationsform der Zukunft für die Informationsversorgung und Informationsbereitstellung? In: B.I.T. online 1999/2000: <http://www.b-i-t-online.de/archiv/1999-02/fachbeit/rushfeja/artikel.htm> (Teil 1)
<http://www.b-i-t-online.de/archiv/1999-03/fach3.htm> (Teil 2)
<http://www.b-i-t-online.de/archiv/1999-04/fach4.htm> (Teil 3)
<http://www.b-i-t-online.de/archiv/2000-01/fach3.htm> (Teil 4)
<http://www.b-i-t-online.de/archiv/2000-02/fach4.htm> (Teil 5)
 (Zugriffe: 22.1.2003)
- Salton, G.; McGill, M.J.: Information Retrieval – Grundlegendes für Informationswissenschaftler. Hamburg, New York, 1987
- Schwarz, I.; Umstätter, W.: Die vernachlässigten Aspekte des Thesaurus: dokumentarische, pragmatische, semantische und syntaktische Einblicke. In: nfd. Information – Wissenschaft und Praxis 50 (4), 1999, S. 197–203 (1999). URL: http://www.ib.hu-berlin.de/~wumsta/infopub/pub1996f/thesaurus_Semiotik.pdf
- Soergel, D.: SemWeb. Proposal for an open, multifunctional, multilingual system for integrated access to knowledge about concepts and terminology. 1996. URL: <http://www.clis.umd.edu/faculty/soergel/soergelsemwebfl.pdf> (Zugriff: 28.4.2003)
- Soergel, D.: The Rise of Ontologies or the Reinvention of Classification. In: Journal of the American Society for Information Science 50 (12), 1999, S. 1119–1120

- Stuckenschmidt, H.; van Harmelen, F.: Ontology-Based Metadata Generation from Semi-Structured Information. In: Proceedings of the International Conference on Knowledge Capture. New York, 2001
- SUO: (Standard Upper Ontology) <http://suo.ieee.org/scopeAndPurpose.html> (Zugriff: 16.3.2003)
- UMLS Applications: <http://www.nlm.nih.gov/research/umls/umlsapps.html> (Zugriff: 22.4.2003)
- UMLS Fact Sheet Metathesaurus:
<http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html> (Zugriff: 16.3.2003)
- UMLS Fact Sheet Semantic Network:
<http://www.nlm.nih.gov/pubs/factsheets/umlsseemn.html> (Zugriff: 16.3.2003)
- UMLS Fact Sheet Specialist Lexicon:
<http://www.nlm.nih.gov/pubs/factsheets/umlslex.html> (Zugriff: 16.3.2003)
- UMLS Fact Sheet: <http://www.nlm.nih.gov/pubs/factsheets/umls.html> (Zugriff: 16.3.2003)
- Uschold, M.; Gruninger, M.: Ontologies: principles, methods, and applications. In: The knowledge Engineering Review 11 (2), 1996, S. 93–136
- Vickery, B. C.: Ontologies. In: Journal of Information Science, 23 (4), 1997, S. 277–286
- Weinstein, P.C.; Birmingham, W.P.: Creating ontological metadata for digital library content and services. In: International Journal on Digital Libraries 2, 1998, S.: 20–37
- Welty, Chr.A.; Jenkins, J.: Formal ontology for subject. In: Data & Knowledge Engineering 31, 1999, S. 155–181